TECHNICAL REPORT
R-98
First version: May, 2023
Last version: Oct, 2023

Causal discovery from observational and interventional data across multiple environments

Adam Li
Department of Computer Science, Columbia University
adam.li@columbia.edu

Amin Jaber
Department of Computer Science, Purdue University
jaber0@purdue.edu

Elias Bareinboim
Department of Computer Science, Columbia University
eb@cs.columbia.edu

Abstract

A fundamental problem in many sciences is the learning of causal structure underlying a system, typically through observation and experimentation. Commonly, one even collects data across multiple domains, such as gene sequencing from different labs, or neural recordings from different species. Although there exist methods for learning the equivalence class of causal diagrams from observational and experimental data, they are meant to operate in a single domain. In this paper, we develop a fundamental approach to structure learning in non-Markovian systems (i.e. when there exist latent confounders) leveraging observational and interventional data collected from multiple domains. Specifically, we start by showing that learning from observational data in multiple domains is equivalent to learning from interventional data with unknown targets in a single domain. But there are also subtleties when considering observational and experimental data. Using causal invariances derived from do-calculus, we define a property called S-Markov that connects interventional distributions from multiple-domains to graphical criterion on a selection diagram. Leveraging the S-Markov property, we introduce a new constraint-based causal discovery algorithm, S-FCI, that can learn from observational and interventional data from different domains. We prove that the algorithm is sound and subsumes existing constraint-based causal discovery algorithms.

1 Introduction

Causal discovery is the process of learning cause-and-effect relationships between variables in a given system, which is many times the final goal of the data scientist or a necessary step towards a more refined causal analysis [1, 2]. The learning process typically leverages constraints from data to infer the corresponding causal diagram. However, it is common that the data constraints do not uniquely identify the full diagram. Therefore, the target of analysis is often an equivalence class (EC) of causal diagrams that encodes constraints found in the data (implied by the underlying unknown causal system).

An EC encodes invariances in the form of graphical constraints, and thus is used to represent all causal diagrams that encode those constraints and invariances. Formal characterizations of ECs are

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

important to understand the output of a learning algorithm and how it relates to the underlying causal system the scientist aims to explain.

ECs are defined with respect to distributional invariances which are implied by the structure of the graph. For example, conditional independences (CI) are implied by d-separations in the causal graph. Hence, it is desirable to formally characterize the EC in the general setting where we have interventional data from multiple domains. A complete graphical characterization would enable i) an efficient representation of the distributional invariances in the data and ii) the ability to translate these data-invariances to graphical constraints (e.g. d-separation).

An early example of an EC when only observational data is available in a single domain is the Markov equivalence class (MEC). The MEC characterizes causal diagrams with the same set of d-separation statements over observed nodes [2–5]. Given interventional (i.e. experimental) data, one can reduce the size of the equivalence class [6, 7]. In the case of known interventional targets, the EC is known as the $\mathcal{I}$-MEC [7–9] and in the case of unknown targets, it is called the Ψ-MEC [6].

In prior research, domain-changes and interventions were treated similarly [10–14]. Nevertheless, various examples across scientific disciplines highlight their distinction (see Table 1). For instance, when extrapolating data-driven conclusions from bonobos to humans, consider Figure 1(b). Notably, the environment/domain, represented by the S-node pointing to X, illustrates differences in kidney function between the species. When applying a CRISPR intervention to a gene linked to kidney protein production (X), researchers investigate the impact of medication (Y) on fluid balance in the body (Z). This intervention is explicitly different from the kidney-function differences between bonobos and humans because the change-in-domain is there regardless of whether or not an intervention is made. This differentiation between interventions and domains holds significance, especially in causal discovery. By leveraging invariances across observational and interventional data from both bonobos and humans, one can learn additional causal relationships. Moreover, conflating these qualitatively distinct settings is generally invalid, as pointed out in transportability analysis [15]. Environmental differences persist regardless of interventions like CRISPR, and kidney function interventions vary between species. Pearl and Bareinboim (2011) introduced clear semantics for S-nodes (environments), offering a unified representation.

In this paper, we investigate structure learning when mixtures of observational and interventional data (known and unknown targets) across multiple domains are available. The multi-domain setting has been analyzed from the lens of selection diagrams, where selection nodes (or S-nodes) encode distributional changes in the mechanisms, or exogenous variables due to a change in domain [16–18]. We will show in this paper a characterization of the EC for selection diagrams. Generalizing the structure learning setting to multiple domains requires a formal treatment because it is a common scenario in the sciences [19–31]; see Table 1 for an example of different settings and related literature). For example, in single-cell sequencing analysis, scientists are interested in analyzing the causal effects of proteins on one another. However, they may typically collect observational and/or experimental data from multiple labs (i.e. domains) and wish to combine them into one dataset. Also, scientists may collect observational and experimental data over multiple species in order to learn more about one specific species, or the relationships among species [25, 27, 32].

The celebrated FCI algorithm and its variants learn a partial ancestral graph (PAG), an MEC of causal diagrams, given purely observational data [1, 2, 33]. The $\mathcal{I}$-FCI (with known targets) and Ψ-FCI (with unknown targets) generalize these results to interventional data, and further reduce the size of the EC to an $\mathcal{I}$-PAG and Ψ-PAG, respectively [6, 7]. However, these algorithms operate in a single domain, or environment and do not account for combining known/unknown target interventions.

Various approaches have been proposed throughout the literature for causal discovery from multiple domains. The works in [10, 13, 34–38] assume Markovianity, a functional model (e.g. linearity) holds, and/or do not take into account arbitrary combinations of observational and interventional data with known and unknown targets. Alternatively, JCI pools data together and performs learning on the combined dataset [14]. Pooling data is an incomplete procedure when considering interventional data within a single domain let alone multiple domains [6][Appendix D.2].

In this paper, we take a principled approach to the multi-domain structure learning problem and formally characterize S-PAGs, the object of learning. This paper introduces the selection-diagram FCI algorithm (S-FCI) that learns from a mixture of observational and interventional data from

Domain	Obs.	Interv. $\mathcal{K}$	Interv. $\mathcal{U}$	Property	FCI-variant	Related Lit.
1	✓	x	x	Markov [39]	[2, 33, 40–43]	[30, 31]
1	✓	✓	x	I-Markov [7, 44]	[7, 8, 44]	[22, 30]
1	✓	x	✓	Ψ-Markov [6]	[6, 13, 45]	[22, 27, 46]
k	✓	x	x	Ψ-Markov (Thm. 1)	[6] (Cor. 5)	[20, 21, 23, 24, 47, 48]
k	✓	✓	✓	S-Markov (Thm. 2)	S-FCI (Thm. 3)	[20–25, 28–31, 46–50]

Table 1: Summary of Markov property results, and related algorithms that learn the ancestral graph based on number of domains and types of interventional (interv.) data provided such as observational (obs.), and known ($\mathcal{K}$) and unknown ($\mathcal{U}$) targets. The last column indicates a brief survey of different fields in ecology, economics, genomics, neurosciences, neurology and medicine that attempt to answer questions at each level. The rows highlighted in "red" are new concepts.

multiple domains to construct an EC of selection diagrams, an S-PAG. Specifically, we contribute the following:

1. **Generalization of standard Markov properties** - We introduce the S-Markov property, which extends and generalizes the normal Markov, I-Markov, and Ψ-Markov properties to the setting of multiple domains with arbitrary mixtures of observational and interventional data with known and unknown targets.
2. **Learning algorithm** - We develop a sound learning algorithm for learning an ~~Markov~~ equivalence class of selection diagrams with observational and/or interventional data across different domains.[1]

1.1 Preliminaries and Notation

Uppercase letters (X) represent random variables, lowercase letters (x) signify assignments, and bold ones ($\mathbf{X}$) indicate sets. The CI relation $\mathbf{X}$ being independent of $\mathbf{Y}$ given $\mathbf{Z}$ is denoted as $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$. The d-separation (or m-separation) of $\mathbf{X}$ from $\mathbf{Y}$ given $\mathbf{Z}$ in graph G is expressed as $(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})_G$. $G_{\overline{X}}$ depicts G with incoming edges to X removed, while $G_{\underline{X}}$ omits all edges outgoing from X. Conventionally, every variable is d-separated from the empty set, denoted as $(X \perp \{\})_G$. Superscripts and subscripts will be dropped where feasible to simplify notation.

Causal Bayesian Network (CBN): Let $P(\mathbf{V})$ be a probability distribution over a set of variables $\mathbf{V}$, and $P_{\mathbf{x}}(\mathbf{V})$ denote the distribution resulting from the *hard intervention* $do(\mathbf{X} = \mathbf{x})$, which sets $\mathbf{X} \subseteq \mathbf{V}$ to constants $\mathbf{x}$. Let $\mathbf{P}^*$ denote the set of all interventional distributions $P_{\mathbf{x}}(\mathbf{V})$, for all $\mathbf{X} \subseteq \mathbf{V}$, including $P(\mathbf{V})$. A directed acyclic graph (DAG) over $\mathbf{V}$ is said to be a *causal Bayesian network* compatible with $\mathbf{P}^*$ if and only if, for all $\mathbf{X} \subseteq \mathbf{V}$, $P_{\mathbf{x}}(\mathbf{v}) = \prod_{\{i|V_i \notin \mathbf{X}\}} P(v_i|\mathbf{pa}_i)$, for all $\mathbf{v}$ consistent with $\mathbf{x}$, and where $\mathbf{Pa}_i$ is the set of parents of V_i [41, 51, pp. 24]. Given that a subset of the variables are unmeasured or latent, $G(\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ will represent the causal graph where $\mathbf{V}$ and $\mathbf{L}$ denote the measured and latent variables, respectively, and $\mathbf{E}$ denotes the edges. Following the convention in [41], for simplicity, a dashed bi-directed edge is used instead of the corresponding latent variables. CI relations can be read from the graph using a graphical criterion known as *d-separation*.

Soft Interventions: Under this type of interventions, the original conditional distributions of the intervened variables $\mathbf{X}$ are replaced with new ones, without completely eliminating the causal effect of the parents. Accordingly, the interventional distribution $P_{\mathbf{X}}(\mathbf{v})$ for $\mathbf{X} \subseteq \mathbf{V}$ is such that $P^*(X_i|Pa_i) \neq P(X_i|Pa_i), \forall X_i \in \mathbf{X}$, and factorizes as follows:

$$P_{\mathbf{X}}(\mathbf{v}) = \sum_{\mathbf{L}} \prod_{\{i|X_i \in \mathbf{X}\}} P^*(x_i|\mathbf{pa}_i) \prod_{\{j|T_j \notin \mathbf{X}\}} P(t_j|\mathbf{pa}_j) \tag{1}$$

In this work, we assume no selection bias and solely consider soft interventions. In the presence of multiple domains, a selection diagram captures commonalities and differences between domains [16, 52, 53]. Represented as $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$, it extends a causal diagram by incorporating S-nodes and their edges. $\binom{N}{2}$ S-nodes, $S^{i,j}$, indicate distribution changes across pairs among N domains, by pointing to nodes in $\mathbf{V}$ whose mechanism is altered between domains i and j. An example

[1] Our algorithm is implemented in open-source MIT-Licensed https://github.com/py-why/dodiscover.

is shown in Figure 1(a), where the S-node is pointing to X, indicating that the distribution of X changes, or that of the latent variable of X is different across the two domains.[2] Similarly, "F-nodes" are auxiliary nodes used in [1, 7, 54] to represent invariances with respect to interventions within the same domain. F-nodes in this paper when written as $F_X^{i,j}$ means it intervenes on X and compares distributions from domains i and j. F_X^i means it compares distributions within domain i. Unlike interventions, domain-shifts potentially alter latent variable distributions or functional relationships and persist irrespective of whether or not external intervention occurs. Distinguishing these concepts enables S-node learning, vital for transportability analysis on ancestral graphs. Appendix Section D.4 elaborates on our distinctions from previous work [11, 13, 14, 36].

Let $\mathbf{S} = \{S^{1,2}, S^{1,3}, ..., S^{N-1,N}\}$ represent $\binom{N}{2}$ S-nodes for distribution changes across domain pairs. When $i = j$, $S^{i,j} = \phi$, indicating there is no S-node for a single domain.

Multi-domain setup The following objects are utilized repeatedly, and introduced here. Our notation borrows from [6] and the transportability literature [55].

1. **Domains**: $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$ denotes a set of N domains.
2. **Intervention targets**: $\mathbf{\Psi}^{\Pi} = \langle \Psi_1^1, \Psi_2^1, ..., \Psi_M^N \rangle$ is an ordered tuple of sets of intervention targets, with different sets of intervention targets occurring within each of the N domains for a total of M intervention target sets. We will denote $\mathbf{\Psi}^i$ as the intervention targets associated with domain i.
3. **Distributions**: $\mathbf{P}^{\Pi} = \langle P_1^1, P_2^1 ..., P_M^N \rangle$ is an ordered tuple of probability distributions that are available to learn from. Denote $\mathbf{P}^i$ as the distributions associated with domain i. There is a one-to-one correspondence between $\mathbf{P}$ and $\mathbf{\Psi}$, such that P_j^i is the distribution associated with targets $\mathbf{\Psi}_j^i$ in domain i.
4. **Known target indices**: $\mathcal{K}$ is a vector of 1's and 0's indicating, which sets of interventions are known-targets. $\mathcal{U} := 1 - \mathcal{K}$ represents therefore an index vector selecting the distributions and interventions with unknown targets. $\mathbf{P}_{\mathcal{K}}$ and $\mathbf{\Psi}_{\mathcal{K}}$ denotes the set of distributions and intervention targets corresponding to the known target interventions.
5. **Causal diagram**: $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, is a shared diagram over the N domains.
6. **Selection diagram**: $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$, extend G with the corresponding S-nodes and their edges to represent each pair of domains. Let $V_{S^{i,j}}$ denote the set of nodes that S-node $S^{i,j}$ points to and $\mathbf{V_S}$ as the set of children for all S-nodes of G_S.

$\mathbf{X}^i$ denotes the ith domain set of variables $\mathbf{X}$, and $X_i \in \mathbf{X}$ indicates the ith variable within $\mathbf{X}$. When discussing intervention targets, $X_j^{i,(k)}$ refers to the jth variable with the kth mechanism change in domain i. For instance, $X^{i,(k)}, X^{i,(l)}$ represent two interventions with distinct mechanisms (k and l) on variable X in domain i. $\{\}^i \in \mathbf{\Psi}$ explicitly denotes the observational distribution for domain i and is by convention a "known-target". For concreteness, say $\mathbf{\Pi} = \{\Pi^1, \Pi^2, \Pi^3\}$ with $\mathbf{P} = \langle P_1^1, P_2^1, P_3^1, P_1^3 \rangle$, $\mathbf{\Psi} = \langle \{\}^1, \{X^{(a)}\}^1, \{X, Y\}^1, \{\}^3 \rangle$, and $\mathcal{K} = [1, 1, 0, 1]$. In words, there are three distributions available in domain 1: P_1^1 is observational, P_2^1 is known-target on X with a specific mechanism change and P_3^1 is unknown-target that intervenes on X and Y simultaneously. In domain 3, P_1^3 is observational. There are no distributions for domain 2.

2 Multi-domain Markov Equivalence Class

Before designing a learning algorithm, one must characterize what can be learned from the given causal graph and its corresponding selection diagram. This section explores ECs in a multi-domain setting with arbitrary mixtures of observational and interventional data. The following assumptions are made throughout the main paper:

Assumption (Shared causal structure)**.** We assume that each environment shares the same causal diagram. That is the S-nodes do not change the underlying causal diagram. □

[2] In the original selection diagram, each S-node points to a single node. Our adaptation simplifies it to a single S-node with multiple connections. Theoretical properties remain unaffected, as shown in the appendix.

This means that the S-nodes do not represent structural changes such as when V_i has a different parent set across domains [3].

Assumption (Observational data is present across domains)**.** We make the simplifying assumption that $\{\} \in \mathbf{\Psi}^i,\ \forall i \in [N]$, that is observational data is present in all domains.

This is a realistic assumption in many scientific applications highlighted in Table 1 [4]. Another assumption we make is that all soft interventions across domains are *distinct*.

Assumption (Distinct interventions across domains)**.** Assume that all interventions across different domains have different mechanisms. That is if $X \in \mathbf{I}^1$ and $X \in \mathbf{J}^2$, X is intervened with a different mechanism. Notationally, we would write $\mathbf{I}^1 = \{X^i\}^1$ and $\mathbf{J}^2 = \{X^j\}^2$.

This is a realistic assumption that precludes the possibility that any interventions that occur in different domains result in the same exact mechanism. For example, even if medication is given to humans and bonobos, it is unrealistic to expect the intervention has the same mechanism of action in each domain. Next, we define an important operation when comparing two different intervention sets.

Definition 2.1 (Symmetrical Difference Operator Δ in Multiple Domains)**.** For domain i and j, given two sets of variables, $\mathbf{I}^i$ and $\mathbf{J}^j$, let $\mathbf{I}^i \Delta \mathbf{J}^j$ denote the symmetrical difference set such that $\mathbf{I}^i \Delta \mathbf{J}^j = \{v \in \mathbf{I}^i | v \notin \mathbf{J}^j\}$ if $V \in \mathbf{I}^i$ and $V \notin \mathbf{J}^j$ or vice versa. □

This operation will identify sets of variables with unique interventional mechanisms across two interventional targets and also track the domain ids. For example, $\mathbf{I}^1 = \{X^1, Y, Z\}^1$ and $\mathbf{J}^2 = \{X^2, Y\}^2$. $\mathbf{I}^1 \Delta \mathbf{J}^2 = \{X, Z\}^{1,2}$. Since selection diagrams are defined with respect to a pair of domains, the causal graph we are actually interested in is a composition of all pairwise selection diagrams between all combinations of domains. In this paper, unless explicitly stated that intervention mechanisms are the same, it is assumed that they are different. For example, consider an intervention on X in domain 1 and paired intervention on X and Y in domain 2. The intervention on X occurs with different mechanism in both domains. I.e. $\mathbf{I}^1 = \{X\} \cup \{Y^1\} = \{X, Y^1\}^1$ and $\mathbf{J}^2 = \{X, Y\}^2$, then $\mathbf{I}^1 \Delta \mathbf{J}^2 = \{X, Y\}^{1,2}$. Denote $\{\}^1 \Delta \{\}^2 = \{\}$. For more details and discussion on the assumptions, see the Appendix.

2.1 Multi-distributional invariances: interventions and change-of-domain

This section elaborates on exactly what type of distributional invariances we characterize in the so called S-Markov EC (see Section 2).

When given only observational data, the celebrated FCI algorithm uses invariances of the form $P(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = P(\mathbf{Y}|\mathbf{X})$ within the same probability distribution, $P(\mathbf{V})$ to characterize the Markov EC [2]. These invariances, or CI statements can be mapped to d-separation statements on the graphical model. The result is the PAG, which is the EC when only observational data is given within a single domain and distribution.

The works in [6–8, 44] build upon the Markov EC to characterize the so called **I**-Markov EC, which uses distributional invariances of the form $P_{\mathbf{W}}(\mathbf{Y}|\mathbf{X}) = P_{\mathbf{W}}(\mathbf{Y}|\mathbf{X})$. $P_{\mathbf{W}}(\mathbf{V})$ is the distribution of $\mathbf{V}$ under some intervention on variables $\mathbf{W}$. Note $\mathbf{W} = \phi$ would denote the observational distribution. Importantly, this sort of invariance is markedly different from that of the CI statements when only observational data is present because one is now comparing probabilities across *different distributions*. These distributional invariances can be characterized graphically by the d-separation property when using an augmented graph with "F-nodes" serving as graphical representations of the differences in distributions due to interventions.

Within this work, we then further generalize the invariances one can consider to establish the S-Markov EC, introduced in 2. These analyze distributional invariances of the form $P^i_W(\mathbf{Y}|\mathbf{X}) = P^j_K(\mathbf{Y}|\mathbf{X})$. Note now the distributions can stem either from a different domain $i \neq j$, or a different intervention set, $W \neq K$. If $P^i_W(\mathbf{Y}|\mathbf{X}) = P^j_K(\mathbf{Y}|\mathbf{X})$, then this means the distribution of $\mathbf{Y}|\mathbf{X}$ is invariant across domains i and j with interventions on $W, K \subseteq V$. When $i = j$, these invariances reduce to

[3] The assumption that there are no structural changes between domains can be relaxed in the context of inference, as specified in [16]. We do not explore this relaxation here in the context of structure learning.

[4] If one can collect experimental data in a domain, it is reasonable that they can also collect observational data. We discuss this further in the Appendix.

the ones considered in the interventional Markov EC. From this perspective, it is clear that multi-domain invariances generalize the invariances analyzed in observational and interventional data in a single-domain.

2.2 S-Markov Property

Now, we are ready to generalize the standard Markov properties [2, 5–7, 41, 56] to the case when observational, and known/unknown-target interventional distributions in multiple domains are available.

Definition 2.2 (S-Markov Property)**.** Consider the multi-domain setup in 1.1. For a fixed $\mathcal{K}$, we say $\mathbf{P}^{\Pi}$, satisfies the S-Markov property with respect to the tuple $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$ if the following holds for disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{Z} \subseteq V$:

1. (Conditional Independences) - For each domain, $\Pi^i \in \mathbf{\Pi}$, and intervention target set, $\mathbf{\Psi}^i_j \in \mathbf{\Psi}^{\Pi}$: $P^i_j(y|w,z) = P^i_j(y|w)$ if $(\mathbf{Y} \perp \mathbf{Z}|\mathbf{W}, \mathbf{S})_{G_S}$
2. (Conditional Invariances) - For each $\Pi^i, \Pi^j \in \mathbf{\Pi}$ and $\mathbf{\Psi}^i_k, \mathbf{\Psi}^j_l \in \mathbf{\Psi}^i$, $P^i_k(y|w) = P^j_l(y|w)$ if $(Y \perp K|W \backslash W_K)_{G_{S_{\overline{W_K}, \overline{R(W)}}}}$, where $\mathbf{K} = (\mathbf{\Psi}^i_k \Delta \mathbf{\Psi}^j_l) \cup \{S^{i,j}\}$, $\mathbf{W_K} = \mathbf{W} \cap \mathbf{K}, \mathbf{R} = \mathbf{K} \backslash \mathbf{W_K}$ and $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ are non-ancestors of $\mathbf{W}$ in G_S.

Denote $S^{\Pi}_{\mathcal{K}}(G_S)$ as the set of distribution tuples that satisfy the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$. □

When there is only a single domain, $\mathbf{\Pi} = \{\Pi^1\}$, the first constraint reduces to standard d-separation on a causal diagram. The second condition is a generalization of the Ψ-Markov property characterization [6], extending conditional invariances to multiple domains. Note the characterization applies to a given causal graph and its $\mathbf{V_S}$ nodes. A selection diagram provides $\mathbf{V_S}$ and the causal graph.

Example 1. Consider the selection diagram in Figure 1(a) with two domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$. Let $\mathbf{P} = \langle P^1_1, P^1_2, P^2_1 \rangle$ be the result of the interventions $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$, $\mathbf{S} = \{S^{1,2}_x\}$ be the set of S-nodes and $\mathcal{K} = [1, 0, 1]$. Since $(Y \perp S^{1,1}|W)_G$ always return True by convention, the second constraint is not applicable when comparing P^1_1, P^1_2. Moreover, comparing $P^1_1(y|x)$ and $P^1_2(y|x)$, note $(Y \perp X)_{G_{\overline{X}}}$ does not hold, so $P^1_1(y|x) = P^1_2(y|x)$ is not required. The same holds when comparing P^2_1 vs P^1_1 and P^2_1 vs P^1_2. Therefore, $\mathbf{P}$ satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$. □

Example 2. Consider the 3-tuple $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$, $\mathcal{K}$ and $\mathbf{P}^{\Pi}$ from Ex 1. Now let $\mathbf{\Psi'}^{\Pi} = \langle \{\}^1, \{Y\}^1 \{\}^2 \rangle$. $K = (\{\}^1 \Delta \{Y\}^1) \cup \{\} = \{Y\}^1$. In this setting, $(X \perp Y)_{G_{\overline{Y}}}$ implies the invariance $P^1_1(X) = P^1_2(X)$, but P^1_2 was generated from an intervention on X and the invariance is not satisfied in $\mathbf{P}$. Therefore $\mathbf{P}$ does not satisfy the S-Markov property with respect to $\langle G, \mathbf{\Psi'}^{\Pi}, \mathbf{V_S} \rangle$. □

The S-Markov property is a generalization of the Markov, I-Markov and Ψ-Markov properties as summarized in Table 1. When there is a single domain, the S-Markov property simplifies to the standard Ψ-Markov or I-Markov property when comparing distributions associated with unknown, or known intervention targets respectively.

Lemma 1 (S-Markov property generalizes the Ψ-Markov property)**.** Consider the multi-domain setup in 1.1 with $\mathbf{\Pi} = \{\Pi^1\}$. If $\mathcal{K}$ is 0 for all non-observational interventions, if $\mathbf{P}^{\Pi}$ satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$, then it also satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi} \rangle$. □

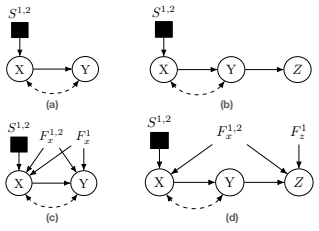

Figure 1: Example selection diagrams (a,b) and their respective augmented graphs (c,d).

Due to space constraints, all the proofs are provided in the Appendix. Consider a few examples stemming from Figure 1.

Example 3 (Markov vs S-Markov property)**.** Let G_S be the selection diagram in Figure 1(b). For an arbitrary set of interventions set $\mathbf{\Psi}^{\Pi}$ and domains $\mathbf{\Pi}$, we have that $(X \perp Z|Y)_G$ implies that $P^i_j(Z|Y,X) = P^i_j(Z|Y)$ for all $\Pi^i \in \mathbf{\Pi}$ and distributions. Thus, the S-Markov property includes the Markov property invariance. However, the Markov property does not capture other invariances that are presented in Def. 2.2. □

Example 4 (Ψ-Markov vs S-Markov property)**.** Let G_S be the selection diagram in Figure 1(b). Let $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{\}^2, \{X\}^1, \{Y\}^1 \rangle$ and $\mathcal{K} = [1, 1, 0, 0]$ for a corresponding $\mathbf{P}^{\Pi}$. The S-Markov property states that there is an invariance $P_1^1(z|y) = P_2^1(z|y) = P_3^1(z|y)$. The I-Markov states the equivalence between $P_1^1(z|y) = P_3^1(z|y)$ and the Ψ-Markov property $P_1^1(z|y) = P_2^1(z|y)$. The S-Markov property captures each of these invariances. □

2.3 Multi-domain observational data

S-nodes introduced through the lens of selection diagrams are augmentations of the causal graph to represent different domains and changes in distributions that may occur [7, 15, 54, 57]. As part of this augmented graph, S-nodes are graphically similar to F-nodes, which have been successfully used to represent interventions [6, 7, 54]. F-nodes are utility nodes that are a parent to each element of the symmetric difference of interventions. They are used to represent invariances between interventional distributions. The significance of these F-nodes will be emphasized in Definition 2.3 and Proposition 1. S-nodes are useful to distinguish, since many causal inference tasks such as transportability rely on knowing the S-node structure [15, 53]. Before developing the full learning algorithm for ECs of selection diagrams, we first study the setting where there is only observational data across different domains. We demonstrate that the S-nodes can be viewed as exactly F-nodes constructed from interventions with unknown targets.

Theorem 1 (Equivalence of Ψ and S Markov property given multi-domain observational distributions)**.** Let G be a shared causal diagram among N domains, $\mathbf{\Pi}$. Let $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ be the corresponding selection diagram. Let $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, ..., \{\}^N \rangle$ and $\mathcal{K} = [1, 1, ..., 1]$, such that for each of the N domains, there is only observational data. Let $\mathbf{P}^{\Pi}$ be an arbitrary set of distributions generated by the corresponding interventions. $\mathbf{P}^{\Pi}$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{V_S} \rangle$, if and only if it satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$. □

Example 5. Let G_S be the selection diagram in Figure 1(b), among two domain $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$. Let $\mathbf{S}^{\Pi} = \{S_x^{1,2}\}$ and $\mathbf{\Psi}^{\Pi}$ and $\mathcal{K}$ be defined with just observational distributions from domains 1 and 2. Consider an arbitrary $\mathbf{P}$ that satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{V_S} \rangle$. This implies the distribution of Z is the same between domains 1 and 2 through the invariance $P^1(Z) = P^2(Z)$. This is the only invariance that is required. Observe that is also the only invariance required by the S-Markov property and thus $\mathbf{P}$ satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$. □

When given observations collected from multiple domains, it is equivalent to collecting distributions with unknown-target interventions. This coincides with other works, which treat different domains and interventions as the same [10, 13]. In this setting, S-nodes have a correspondence to the augmented graph's F-nodes in [6]. However, this simplification is not warranted when we consider interventions that occur in different domains.

2.4 Mixture of multi-domain observational and interventional data

Next, we analyze the general setting with multi-domain observational and interventional data. The S-Markov property in Definition 2.2 may be quite challenging to evaluate in practice, since it involves surgically altering the selection diagram. One can leverage a graphical approach that encodes the symmetric differences of interventions as an F-node [7].

Definition 2.3 (Augmented graph)**.** Consider the multi-domain setup 1.1. Let the multiset $\mathcal{I}$ be defined as such $\mathcal{I} = \{\mathbf{K}_1, \mathbf{K}_2, ...\mathbf{K}_k\} = \{\mathbf{K} | \mathbf{I}^i, \mathbf{J}^j \in \mathbf{\Psi}^{\Pi} \wedge (\mathbf{I}^i \Delta \mathbf{J}^j) \cup \mathbf{V}_{S^{i,j}} = \mathbf{K}\}$. The augmented graph G_S with respect to $\mathbf{\Psi}^{\Pi}$ and $\mathbf{V_S}$ is denoted $Aug_{\Psi, V_S}(G_S)$ and constructed as follows: $Aug_{\Psi, V_S}(G_S) = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \mathcal{F}, \mathbf{E} \cup \mathbf{E_S} \cup \mathcal{E})$, where $\mathcal{F} = \{F_i^{j,k}\}^{j,k \in [N]}$ is the set of added F-nodes and $\mathcal{E} = \{(F_i^{j,k}, l)\}_{l \in \mathbf{K}_i}$ is the set of added F-node edges. Denote $F_k^{i,i} = F_k^i$ as an F-node representing the kth symmetric difference of intervention targets within domain i and $F_k^{i,j}$ as an F-node from comparing targets between domain i and j. □

The F-nodes constructed consider the symmetrical difference between every possible pair of interventions across different domains and also within the same domain, $\mathbf{I}^i \Delta \mathbf{J}^j$. The result is a augmented selection diagram with the original causal structure augmented with F-nodes and their additional edges. The F-nodes are a parent to each node in $\mathbf{K}$, which is constructed by the symmetric difference of the intervention targets and the children of the corresponding S-node (if comparing targets in different domains). For example, Figure 1(c) shows an augmented graph with F-nodes

constructed comparing each distribution. This augmented graph is used to succinctly represent S-Markov equivalence in a graph without graphical mutilations.

Proposition 1 (Graphical S-Markov Property)**.** Consider the multi-domain setup 1.1. Let $Aug_{\Psi,V_S}(G_S)$ be the augmented graph of G_S, where $\mathcal{F}_\mathcal{E} = \{F_i^{j,k}\}$ is the set of added F-nodes. Let $\mathbf{K}_i^{j,k}$ be the set of nodes adjacent to each F-node $F_i^{j,k}$ plus the S-node $S^{j,k}$. The following equivalence relations hold for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:

$$(Y \perp Z|W, \mathbf{S})_{G_S} \iff (Y \perp Z|W, F_\mathcal{E}, \mathbf{S})_{Aug_{\Psi,V_S}(G_S)} \tag{2}$$

$$(Y \perp K_i^{j,k}|W \backslash W_i)_{G_{S\underline{W_i^j},\overline{R(W)}}} \iff (Y \perp \{F_i^{j,k}, S^{j,k}\}|W, \mathcal{F}_\mathcal{E} \backslash F_i^{j,k}, \mathbf{S}_{[N]\backslash j})_{Aug_{\Psi,S}(G_S)} \tag{3}$$

where $\mathbf{W}_i^{j,k} = \mathbf{W} \cap \mathbf{K}_i^{j,k}, \mathbf{R} = \mathbf{K}_i^{j,k} \backslash \mathbf{W}_i^{j,k}$. □

This proposition allows one to map invariances present in the model to d-separation statements on a graph, which provides an efficient representation.

Example 6. Consider the augmented graph in Figure 1(b) with intervention $\mathbf{\Psi} = \langle\{\}^1, \{Z\}^1, \{\}^2\rangle$. By Prop. 1, we can directly test the S-Markov properties on the graph without surgically altering the graph. For example, $(Y \perp Z)_{G_{\overline{Z}}}$ can be tested by $(Y \perp F_z^1, \mathbf{S})_{Aug_{\Psi,V_S}(G_S)}$ to determine if $P_{\{\}}^1(y) = P_z^1(y)$. In addition, we can also test if across-domain distributional invariances should hold. Since $(Y \perp F_z^{1,2}\}|X, Z, \mathbf{S})_{Aug_{\Psi,V_S}(G_S)}$ does not hold, then the invariance $P_z^1(Y|X, Z) = P_{\{\}}^2(Y|X, Z)$ is not required. The S-node's effect is present through the added F-node, $F_z^{1,2}$. □

Maximal ancestral graphs (MAGs) are a compact and convenient way of representing the constraints in augmented graphs represented by d-separation [58]. We formalize the corresponding ancestral graph for causal graphs representing multiple domains in Def. 2.4, which encode the same constraints as the ones presented in the S-Markov property.

Definition 2.4 (S-MAG)**.** Given the multi-domain setup 1.1, a S-MAG is the MAG constructed from $Aug_{\Psi,V_S}(G_S)$. That is MAG$(Aug_{\Psi,V_S}(G_S))$. □

Example 7. Consider the selection diagram induced by the causal diagram and S-node structure in Figure 1(a). Let $\mathbf{\Psi}^\Pi = \langle\{\}^1, \{X\}^1, \{\}^2\rangle$. The $Aug_{\Psi,V_S}(G_S)$ is the same causal diagram with $F_x^1 \to X$, $F_x^{1,2} \to X$. Then the corresponding S-MAG is $MAG(Aug_{\Psi,V_S}(G_S)) = \{X \leftarrow F_x^1 \to Y, X \leftarrow F_x^{1,2} \to Y, X \leftarrow S^{1,2} \to Y, X \to Y\}$. □

Next, we characterize when two S-MAGs are S-Markov equivalent using purely graphical criterion.

Theorem 2 (S-Markov Characterization)**.** Let there be two causal graphs $G^1 = (\mathbf{V} \cup \mathbf{L}_1, \mathbf{E}_1)$, $G^2 = (\mathbf{V} \cup \mathbf{L}_2, \mathbf{E}_2)$ with $G_S^1 = (\mathbf{V} \cup \mathbf{L}_1 \cup \mathbf{S}, \mathbf{E}_1 \cup \mathbf{E}_{\mathbf{S_1}})$ and $G_S^2 = (\mathbf{V} \cup \mathbf{L}_2 \cup \mathbf{S}, \mathbf{E}_1 \cup \mathbf{E}_{\mathbf{S_2}})$ the corresponding selection diagrams and the intervention targets, $\mathbf{\Psi_1}^\Pi$, $\mathbf{\Psi_2}^\Pi$. Let $\mathcal{K}$ be a fixed index vector of known intervention targets that is shared by the two causal diagrams. Assume that the symmetrical difference sets are indexed in both sets in the same pattern such that correspondence between F-nodes and S-nodes are the same in M_1 and M_2. Then $\langle G^1, \mathbf{\Psi_1}^\Pi, \mathbf{V_{S_1}}\rangle$ and $\langle G^2, \mathbf{\Psi_2}^\Pi, \mathbf{V_{S_2}}\rangle$ are S-Markov equivalent if and only if for $M_1 = MAG(Aug_{\mathbf{\Psi_1},\mathbf{V_S}}(G_S^1))$ and $M_2 = MAG(Aug_{\mathbf{\Psi_2},\mathbf{V_S}}(G_S^2))$:

1. M_1 and M_2 have the same skeleton
2. M_1 and M_2 have the same unshielded colliders
3. If a path p is a discriminating path for a node Y in both M_1 and M_2, then Y is a collider on the path in one graph if and only if it is a collider on the path in the other. □

Thm. 2 provides a graphical criterion for comparing now two sets of causal diagrams, intervention targets, and S-node structure to determine if they are S-Markov equivalent.

Example 8. Consider the triplets $\langle G, \mathbf{\Psi}^\Pi, \mathbf{V_S}\rangle$ from Ex. 1 and $\langle G, \mathbf{\Psi'}^\Pi, \mathbf{V_S}\rangle$ from Ex. 2. In $M_1 = MAG(Aug_{\Psi',V_S}(G))$, the F-node F_x^1 from intervening on X will be adjacent to both X and Y due to the inducing path. However, in $M_2 = MAG(Aug_{\Psi,V_S}(G))$, the F-node F_y^1 from intervening on Y will be adjacent to only Y. Therefore, M_1 and M_2 have differing skeletons and thus are not S-Markov equivalent. □

As a result of this characterization, we can now turn our attention to learning the actual graphical structure.

Algorithm 1 S-FCI: Algorithm for Learning a S-PAG - $SepSet$ the separating sets, $\mathbf{S}$ is the S-node set, $\mathcal{F}^{\Pi}$ the F-node set, $\mathbf{H}$ maps each pair of known-targets symmetric diffs., and σ maps each pair of distributions to a pair of domains.

Input: Tuple of distributions $\mathbf{P}^{\Pi} = \langle P_1^1, ..., P_m^N \rangle$, vector of known intervention targets $\mathcal{K}$ and $\mathbf{\Psi}^{\Pi}$.
Output: S-PAG, $\mathcal{P}$
$\mathbf{S}, \mathcal{F} \leftarrow \phi, k \leftarrow 0, \sigma : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}, \mathbf{H} \leftarrow \phi$
$(\mathbf{S}, \mathcal{F}, \mathbf{H}, \sigma) \leftarrow$ CreateAugmentedNodes($\mathbf{\Psi}^{\Pi}, V$) (see Alg. D.2)
Phase I: Learn skeleton
for all pairs $X, Y \in \mathbf{V} \cup \mathcal{F} \cup \mathbf{S}$ **do**
$SepSet(X, Y), SepFlag \leftarrow$ GeneralizedDoConstraints($X, Y, \mathcal{F}, \mathbf{S}, \sigma, \mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V}$) (see Alg. D.4)
if SepFlag = True **then**
Remove edge between X and Y
Phase IIa: Orient unshielded colliders
For every unshielded triple $\langle X, Y, Z \rangle$ in $\mathcal{P}$ orient it as a collider iff $Z \notin SepSet(X, Y)$
Phase IIb: Apply logical orientation rules
R1-7: Apply 7 FCI rules from [39] and following two rules until none apply.
Rule 8': For $F_k^{i,j} \in \mathcal{F}^{\Pi}$ and for $S^{i,j} \in \mathbf{S}$, orient adjacent edges out of $F_k^{i,j}$ and $S^{i,j}$.
Rule 9': For $F_k^{i,j} \in \mathcal{F}^{\Pi}$ with $X \in H_k^{i,j}$, that is adjacent to a node $Y \notin H_k^{i,j}$, if $|H_k^{i,j}| = 1$, then orient $X \rightarrow Y$.

3 Causal Discovery From Multiple Domains

We investigate in this section how to learn a EC from a mixture of observational and interventional data generated from multiple domains $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$. The graphical characterization of the $S-MAG$ object and the equivalent graphical characterization in Thm. 2 motivates us to define the $S-PAG$.

Definition 3.1 (S-PAG)**.** Consider the multi-domain setup 1.1. Let $M = MAG(Aug_{\Psi, V_S}(G))$ and let $[M]$ be the set of S-MAGs corresponding to all the triplets $\langle G', \mathbf{\Psi'}^{\Pi}, \mathbf{V_S} \rangle$ that are S-Markov equivalent to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$. The S-PAG for $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$, denoted $\mathcal{P}$ is a graph such that:

1. $\mathcal{P}$ has the same adjacencies as M and any member of [M] does and
2. every non-circle mark (tail or arrowhead) in $\mathcal{P}$ is an invariant mark in [M] (i.e. present in all S-MAGs). □

The S-PAG is a valid PAG by construction, and it generalizes PAG and Ψ-PAGs in the single-domain setting [6, 42]. The F-nodes are not so much "random variables" as they are graphical models that represent different domains and interventional distributions in this equivalence class. Next, we introduce a generalization of c-faithfulness [6] that enables causal discovery from multi-domain data.

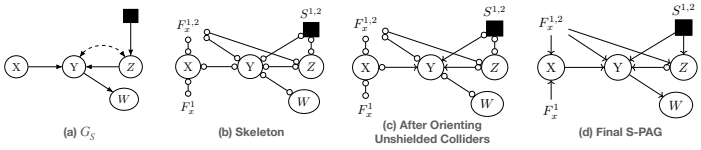

Figure 2: Example of S-FCI applied with $\mathbf{\Psi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$ and $\mathcal{K} = [1, 1, 1]$. The S-node representing domain-shift between domains 1 and 2 is the black square in (a).

Definition 3.2 (S-faithfulness)**.** Consider a causal diagram G and its corresponding selection diagram G_S over N domains. A tuple of distributions $\langle \mathbf{P_I} \rangle_{\mathbf{I} \in \mathbf{\Psi}^{\Pi}} \in S_{\mathcal{K}}^{\Pi}(G)$ is called s-faithful to G_S if the converse of each of the S-Markov conditions (Definition 2.2) holds. □

The new algorithm, called S-FCI is shown in Alg. 1. Due to space constraints, we only include the high-level algorithm here. The algorithm proceeds by first constructing the augmented graph using Alg. D.2, by adding S-nodes and F-nodes to represent every pair of distributions. Then it uses hypothesis testing to learn invariances in the skeleton (Alg. D.3) and finally applies orientation rules (Alg. D.5). S-FCI learns the skeleton by mapping pairs of distributions in $\mathbf{P}^{\Pi}$ to F-nodes, or S-nodes by testing for the distributional invariances discussed in Section 2.1. Def. 2.2 and Prop. 1 connect

these invariances to graphical criterion, which allow us to reconstruct the skeleton of the causal diagram. Interventional distributions across domains are used to learn F-node structure, and whereas observational distributions across domains are used to learn S-node structure. Besides the standard FCI rules that apply in the absence of selection bias, the algorithm also applies the following rules R8'-9'.

Rule 8' (Augmented Node Edges) - We orient edges out of F-nodes.
Rule 9' (Identifiable Inducing Paths) - If $F_k^{i,j} \in \mathcal{F}$ is adjacent to a $Y \notin H_k^{i,j}$ known-target node and we know that the intervention target is node X, one can orient $X \to Y$ because the $F_k^{i,j} \to Y$ is only present due to an inducing path between X and Y.

In Figure 2, the different stages of the S-FCI algorithm are shown. Next we prove the proposed S-FCI algorithm is sound.

Theorem 3 (S-FCI Soundness)**.** Given $\mathcal{K}$, let $\mathbf{P}^{\mathbf{\Pi}}$ be generated by some unknown triplet $\langle G, \mathbf{\Psi}^{\mathbf{\Pi}}, \mathbf{V_S} \rangle$ from domains $\mathbf{\Pi}$ with a corresponding selection diagram G_S and is s-faithful to the selection diagram G_S. S-FCI algorithm is sound (i.e. every adjacency and orientation in $\mathcal{P}_{\text{S-FCI}}$, the S-PAG learned by S-FCI, is common for $MAG(Aug_{\Psi,V_S}(G))$). □

Next, we illustrate some subtleties between the S-FCI and related algorithms that say pool observational and interventional distributions, ignoring the domain change. The example is motivated from biomedical sciences, where interventions are commonly performed in different domains and the goal is to leverage all datasets for learning. A group of scientists are trying to determine the causal structure of a set of proteins, but leverage data across the lab and hospital setting. Different experiments are run in each setting and combined into a single dataset [29]. We provide additional examples and commentary on the S-FCI subtleties in the Appendix.

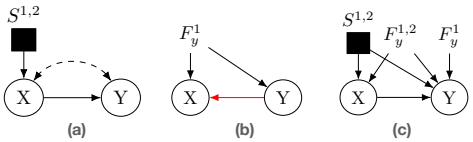

Figure 3: Causal diagrams related to examples 9 - selection diagram with an intervention at Y, and S-node pointing to X (a), the graph after applying unsound rule from $\mathcal{I}$-FCI (b) and the S-PAG learned by S-FCI (c).

Example 9. Let G_S be a selection diagram as shown in Figure 3(a). Let $\mathbf{\Pi} = \langle \Pi^1, \Pi^2 \rangle$ be the set of domains representing the lab (Π^1) and the hospital (Π^2). These are a tuple of distributions $\mathbf{P} = \langle P_1^1, P_1^2 \rangle$ with intervention targets $\mathbf{\Psi}^{\mathbf{\Pi}} = \langle \{\}^1, \{Y\}^1, \{\}^2 \rangle$ and $\mathcal{K} = [1, 1, 1]$, where X represents some protein in the dataset.

In this example, let G_S be the true selection diagram as shown in Figure 3(a). Given the interventional and observational data, we may be tempted to use the $\mathcal{I}$-FCI algorithm and simply pool the observational data, while ignoring the domain differences [7], but this would learn the graph in Figure 3(b) with an incorrect orientation (shown as the red edge). This I-PAG only contains one F-node because there is only two distributions: i) the pooled observational data and ii) the data resulting from intervention on Y. Applying R9 of the $\mathcal{I}$-FCI algorithm incorrectly orients the edge $X \leftarrow Y$. Thus, R9 of the $\mathcal{I}$-FCI algorithm is not sound when the domains are ignored [7, 44].

Figure 3(c) contains what S-FCI would recover. Intuitively, one should learn (c) instead of (b) because even though there is a change in distribution among X and Y, one cannot ascertain whether there is an inducing path from F_y^1 to X, or a change in distribution due to the domain. □

4 Conclusions

In this paper, we introduced a generalized Markov property called S-Markov, which defines a new equivalence class (EC), the S-PAG, representing the constraints found across observational and experimental distributions collected from multiple domains. Building on this new characterization, we develop a causal discovery algorithm called S-FCI, which subsumes FCI, $\mathcal{I}$-FCI and Ψ-FCI, and accepts as input a mixture of observational and interventional data from multiple domains. Future interesting work would involve relaxing the assumptions made in this paper, and leveraging the characterization of the EC for downstream causal ID and estimation tasks.

Acknowledgements

AL was supported by the NSF Computing Innovation Fellowship (#2127309). EB was supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] Judea Pearl. *Causality: Models, reasoning, and inference.* 2nd. Cambridge University Press, 2009.

[2] P. Spirtes, C. Glymour, and R. Scheines. "Causation, Prediction, and Search." In: 81 (1993). Place: New York, NY Publisher: Springer New York.

[3] T. S. Verma and J. Pearl. *An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation*. arXiv:1303.5435 [cs]. 2013.

[4] P. L. Spirtes, C. Meek, and T. S. Richardson. *Causal Inference in the Presence of Latent Variables and Selection Bias*. arXiv:1302.4983 [cs]. 2013.

[5] C. Meek. *Causal Inference and Causal Explanation with Background Knowledge*. arXiv:1302.4972 [cs]. 2013.

[6] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. "Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 9551–9561.

[7] M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. "Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions." In: *Advances in Neural Information Processing Systems* 32 (2019).

[8] A. Hauser and P. Bühlmann. *Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs*. arXiv:1104.2808 [cs, math, stat]. 2012.

[9] A. Hauser and P. Bühlmann. "Two Optimal Strategies for Active Learning of Causal Models from Interventional Data." In: *International Journal of Approximate Reasoning* 55.4 (2014). arXiv:1205.4174 [cs, stat], pp. 926–939.

[10] R. Perry, J. von Kügelgen, and B. Schölkopf. *Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis*. arXiv:2206.02013 [cs, stat]. 2022.

[11] B. Huang, K. Zhang, M. Gong, and C. Glymour. "Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models." en. In: *Proceedings of the 36th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, 2019, pp. 2901–2910.

[12] B. Huang, C. J. H. Low, F. Xie, C. Glymour, and K. Zhang. "Latent hierarchical causal structure discovery with rank constraints." In: *arXiv preprint arXiv:2210.01798* (2022).

[13] J. Peters, P. Bühlmann, and N. Meinshausen. *Causal inference using invariant prediction: identification and confidence intervals*. arXiv:1501.01332 [stat]. 2015.

[14] J. M. Mooij, S. Magliacane, and T. Claassen. "Joint causal inference from multiple contexts." In: *The Journal of Machine Learning Research* 21.1 (2020), 99:3919–99:4026.

[15] J. Pearl and E. Bareinboim. "Transportability of Causal and Statistical Relations: A Formal Approach." en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 25.1 (2011). Number: 1, pp. 247–254.

[16] E. Bareinboim and J. Pearl. "Transportability of Causal Effects: Completeness Results." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 26.1 (2012), pp. 698–704.

[17] J. Pearl and E. Bareinboim. "Transportability across studies: A formal approach." In: (2018).

[18] J. D. Correa and E. Bareinboim. "From Statistical Transportability to Estimating the Effect of Stochastic Interventions." In: ().

[19] T. R. Frieden. "Evidence for Health Decision Making — Beyond Randomized, Controlled Trials." en. In: *New England Journal of Medicine* 377.5 (2017). Ed. by J. M. Drazen, D. P. Harrington, J. J. McMurray, J. H. Ware, and J. Woodcock, pp. 465–475.

[20] A. Li, S. Inati, K. Zaghloul, and S. Sarma. "Fragility in Epileptic Networks : the Epileptogenic Zone." In: 2017, pp. 1–8.

[21] A. Li, C. Huynh, Z. Fitzgerald, I. Cajigas, D. Brusko, J. Jagid, A. O. Claudio, A. M. Kanner, J. Hopp, S. Chen, J. Haagensen, E. Johnson, W. Anderson, N. Crone, S. Inati, K. A. Zaghloul, J. Bulacio, J. Gonzalez-Martinez, and S. V. Sarma. "Neural fragility as an EEG marker of the seizure onset zone." en. In: *Nature Neuroscience* 24.10 (2021). Number: 10 Publisher: Nature Publishing Group, pp. 1465–1474.

[22] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. "Causal protein-signaling networks derived from multiparameter single-cell data." eng. In: *Science (New York, N.Y.)* 308.5721 (2005), pp. 523–529.

[23] J. M. Bernabei, A. Li, A. Y. Revell, R. J. Smith, K. M. Gunnarsdottir, I. Z. Ong, K. A. Davis, N. Sinha, S. Sarma, and B. Litt. "Quantitative approaches to guide epilepsy surgery from intracranial EEG." In: *Brain* (2023), awad007.

[24] A. Palepu, A. Li, Z. Fitzgerald, K. Hu, J. Costacurta, J. Bulacio, J. Martinez-Gonzalez, and S. V. Sarma. "Evaluating Invasive EEG Implantations with Structural Imaging Data and Functional Scalp EEG Recordings from Epilepsy Patients." eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2019 (2019), pp. 3866–3869.

[25] S. Nolte and J. Call. "Targeted helping and cooperation in zoo-living chimpanzees and bonobos." eng. In: *Royal Society Open Science* 8.3 (2021), p. 201688.

[26] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. "Key challenges for delivering clinical impact with artificial intelligence." In: *BMC Medicine* 17.1 (2019), pp. 195–195.

[27] C. D. Stimpson, N. Barger, J. P. Taglialatela, A. Gendron-Fitzpatrick, P. R. Hof, W. D. Hopkins, and C. C. Sherwood. "Differential serotonergic innervation of the amygdala in bonobos and chimpanzees." In: *Social Cognitive and Affective Neuroscience* 11.3 (2016), pp. 413–422.

[28] E. A. Petersen, T. G. Stauss, J. A. Scowcroft, E. S. Brooks, J. L. White, S. M. Sills, K. Amirdelfan, M. N. Guirguis, J. Xu, C. Yu, A. Nairizi, D. G. Patterson, K. C. Tsoulfas, M. J. Creamer, V. Galan, R. H. Bundschu, C. A. Paul, N. D. Mehta, H. Choi, D. Sayed, S. P. Lad, D. J. DiBenedetto, K. A. Sethi, J. H. Goree, M. T. Bennett, N. J. Harrison, A. F. Israel, P. Chang, P. W. Wu, G. Gekht, C. E. Argoff, C. E. Nasr, R. S. Taylor, J. Subbaroyan, B. E. Gliner, D. L. Caraway, and N. A. Mekhail. "Effect of High-frequency (10-kHz) Spinal Cord Stimulation in Patients With Painful Diabetic Neuropathy: A Randomized Clinical Trial." eng. In: *JAMA neurology* 78.6 (2021), pp. 687–698.

[29] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. "Batch effects and the effective design of single-cell gene expression studies." en. In: *Scientific Reports* 7.1 (2017). Number: 1 Publisher: Nature Publishing Group, p. 39921.

[30] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. "Functional discovery via a compendium of expression profiles." eng. In: *Cell* 102.1 (2000), pp. 109–126.

[31] X. Shen, S. Ma, P. Vemuri, and G. Simon. "Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology." en. In: *Scientific Reports* 10.1 (2020). Number: 1 Publisher: Nature Publishing Group, p. 2975.

[32] D. Ehrens, A. Li, F. Aeed, Y. Schiller, and S. V. Sarma. "Network Fragility for Seizure Genesis in an Acute in vivo Model of Epilepsy." eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2020 (2020), pp. 3695–3698.

[33] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. "Learning high-dimensional directed acyclic graphs with latent and selection variables." In: *Annals of Statistics* 40.1 (2011), pp. 294–321.

[34] A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang. *Learning Causal Structures Using Regression Invariance*. arXiv:1705.09644 [cs, stat]. 2017.

[35] C. Heinze-Deml, J. Peters, and N. Meinshausen. *Invariant Causal Prediction for Nonlinear Models*. arXiv:1706.08576 [stat]. 2018.

[36] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. *Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes*. arXiv:1903.01672 [cs, stat]. 2020.

[37] A. Ghassami, N. Kiyavash, B. Huang, and K. Zhang. "Multi-domain Causal Structure Learning in Linear Systems." In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

[38] Y. Zeng, S. Shimizu, R. Cai, F. Xie, M. Yamamoto, and Z. Hao. "Causal Discovery with Multi-Domain LiNGAM for Latent Factors." en. In: *Proceedings of The 2021 Causal Analysis Workshop Series*. ISSN: 2640-3498. PMLR, 2021, pp. 1–4.

[39] J. Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias." In: *Artificial Intelligence* 172.16-17 (2008). Publisher: Elsevier, pp. 1873–1896.

[40] P. Spirtes, C. Glymour, and R. Scheines. "From probability to causality." In: *Philosophical Studies* 64 (1991), pp. 1–36.

[41] J. Pearl and D. Mackenzie. *The book of why : the new science of cause and effect*. Pages: 418. 2019.

[42] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. "Learning high-dimensional directed acyclic graphs with latent and selection variables." In: *The Annals of Statistics* 40.1 (2012). arXiv:1104.5617 [cs, math, stat].

[43] D. Colombo and M. H. Maathuis. "Order-Independent Constraint-Based Causal Structure Learning." In: *Journal of Machine Learning Research* 15 (2014), pp. 3921–3962.

[44] K. D. Yang, A. Katcoff, and C. Uhler. *Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions*. arXiv:1802.06310 [math, stat]. 2019.

[45] D. Eaton and K. Murphy. "Exact Bayesian structure learning from uncertain interventions." en. In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. PMLR, 2007, pp. 107–114.

[46] R. J. Smith, M. A. Hays, G. Kamali, C. Coogan, N. E. Crone, J. Y. Kang, and S. V. Sarma. "Stimulating native seizures with neural resonance: a new approach to localize the seizure onset zone." In: *Brain* 145.11 (2022), pp. 3886–3900.

[47] A. Li, P. Myers, N. Warsi, K. M. Gunnarsdottir, S. Kim, V. Jirsa, A. Ochi, H. Otusbo, G. M. Ibrahim, and S. V. Sarma. *Neural Fragility of the Intracranial EEG Network Decreases after Surgical Resection of the Epileptogenic Zone*. en. Pages: 2021.07.07.21259385. 2022.

[48] K. M. Gunnarsdottir, A. Li, R. J. Smith, J.-Y. Kang, A. Korzeniewska, N. E. Crone, A. G. Rouse, J. J. Cheng, M. J. Kinsman, P. Landazuri, U. Uysal, C. M. Ulloa, N. Cameron, I. Cajigas, J. Jagid, A. Kanner, T. Elarjani, M. M. Bicchi, S. Inati, K. A. Zaghloul, V. L. Boerwinkle, S. Wyckoff, N. Barot, J. Gonzalez-Martinez, and S. V. Sarma. "Source-sink connectivity: a novel interictal EEG marker for seizure localization." In: *Brain* 145.11 (2022), pp. 3901–3915.

[49] K. Jo Black and M. Richards. "Eco-gentrification and who benefits from urban green amenities: NYC's high Line." en. In: *Landscape and Urban Planning* 204 (2020), p. 103900.

[50] A. M. Lozano, N. Lipsman, H. Bergman, P. Brown, S. Chabardes, J. W. Chang, K. Matthews, C. C. McIntyre, T. E. Schlaepfer, M. Schulder, Y. Temel, J. Volkmann, and J. K. Krauss. "Deep brain stimulation: current challenges and future directions." In: *Nature reviews. Neurology* 15.3 (2019), pp. 148–160.

[51] E. Bareinboim and J. Pearl. "Causal Inference by Surrogate Experiments: z-Identifiability." In: *Uncertainty in Artificial Intelligence - Proceedings of the 28th Conference, UAI 2012* (2012), pp. 113–120.

[52] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. "On Pearl's Hierarchy and the Foundations of Causal Inference." In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Vol. 36. New York, NY, USA: Association for Computing Machinery, 2022, pp. 507–556.

[53] E. Bareinboim and J. Pearl. "Meta-Transportability of Causal Effects: A Formal Approach." en. In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. PMLR, 2013, pp. 135–143.

[54] A. P. Dawid. “Influence Diagrams for Causal Modelling and Inference.” In: *International Statistical Review / Revue Internationale de Statistique* 70.2 (2002). Publisher: [Wiley, International Statistical Institute (ISI)], pp. 161–189.

[55] E. Bareinboim and J. Pearl. “Causal inference and the data-fusion problem.” In: *Proceedings of the National Academy of Sciences* 113.27 (2016). Publisher: National Academy of Sciences, pp. 7345–7352.

[56] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[57] J. D. Correa and E. Bareinboim. “General Transportability of Soft Interventions: Completeness Results.” In: ().

[58] T. Richardson and P. Spirtes. “Ancestral graph Markov models.” In: *The Annals of Statistics* 30.4 (2002). Publisher: Institute of Mathematical Statistics, pp. 962–1030.

[59] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. “Independence properties of directed markov fields.” en. In: *Networks* 20.5 (1990). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/net.3230200503, pp. 491–505.

[60] J. Zhang and G. F. Cooper. “Causal Reasoning with Ancestral Graphs.” In: *Journal of Machine Learning Research* 9 (2008), pp. 1437–1474.

[61] J. Zhang. *A Characterization of Markov Equivalence Classes for Directed Acyclic Graphs with Latent Variables*. arXiv:1206.5282 [cs, stat]. 2012.

[62] C. Meek. *Strong Completeness and Faithfulness in Bayesian Networks*. arXiv:1302.4973 [cs]. 2013.

[63] J. Correa and E. Bareinboim. “A Calculus for Stochastic Interventions:Causal Effect Identification and Surrogate Experiments.” en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.06 (2020). Number: 06, pp. 10093–10100.

[64] J. Correa and E. Bareinboim. “General Transportability of Soft Interventions: Completeness Results.” In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 10902–10912.

[65] J. M. Robins, M. A. Hernán, and B. Brumback. “Marginal structural models and causal inference in epidemiology.” eng. In: *Epidemiology (Cambridge, Mass.)* 11.5 (2000), pp. 550–560.

[66] D. Geiger, T. Verma, and J. Pearl. “d-Separation: From Theorems to Algorithms.” en. In: *Machine Intelligence and Pattern Recognition*. Ed. by M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer. Vol. 10. Uncertainty in Artificial Intelligence. North-Holland, 1990, pp. 139–148.

[67] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. “Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning.” In: ().

[68] A. P. Dawid. “Conditional Independence in Statistical Theory.” en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1 (1979). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1979.tb01052.x, pp. 1–15.

[69] J. Pearl. “Causal Diagrams for Empirical Research.” In: *Biometrika* 82.4 (1995). Publisher: [Oxford University Press, Biometrika Trust], pp. 669–688.

[70] A. A. Hagberg, D. A. Schult, and P. J. Swart. “Exploring Network Structure, Dynamics, and Function using NetworkX.” In: *Proceedings of the 7th Python in Science Conference*. Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, 2008, pp. 11–15.

[71] C. Squires, Y. Wang, and C. Uhler. *Permutation-Based Causal Structure Learning with Unknown Intervention Targets*. arXiv:1910.09007 [stat]. 2020.

[72] A. Li, J. Lee, F. Montagna, C. Trevino, and R. Ness. *Dodiscover: Causal discovery algorithms in Python.*

[73] A. Li, J. Lee, and A. Roy. *Pywhy-Graphs: Causal graphs that are networkx-compliant for the py-why ecosystem.*

[74] J. Park, U. Shalit, B. Schölkopf, and K. Muandet. *Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression*. arXiv:2102.08208 [cs, stat]. 2021.

[75] J. M. Mooij and T. Claassen. "Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles." en. In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. ISSN: 2640-3498. PMLR, 2020, pp. 1159–1168.

[76] R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian Networks in R: with Applications in Systems Biology*. en. New York, NY: Springer New York, 2013.

[77] A. Ankan and A. Panda. "pgmpy: Probabilistic graphical models using python." In: *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer, 2015.

[78] P. Hünermund and E. Bareinboim. *Causal Inference and Data Fusion in Econometrics*. arXiv:1912.09104 [econ]. 2023.

[79] D. T. Campbell, J. C. Stanley, and N. L. Gage. *Experimental and quasi-experimental designs for research*. Experimental and quasi-experimental designs for research. Pages: ix, 84. Boston, MA, US: Houghton, Mifflin and Company, 1963.

[80] C. F. Manski. *Identification for Prediction and Decision:* Cambridge, MA: Harvard University Press, 2008.

[81] S. Wasserman. "Review of Statistical Methods for Meta-Analysis." In: *Journal of Educational Statistics* 13.1 (1988). Publisher: [Sage Publications, Inc., American Educational Research Association, American Statistical Association], pp. 75–78.

[82] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage Learning, 2002.

[83] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge: Cambridge University Press, 2007.

Appendix

Contents

D.1 Proofs

Here, we provide the detailed proofs of theoretical results in the main paper. First, we review some fundamental definitions and results that guide the main results

D.1.1 Background results

In this section, we centralize theoretical results in relation to the theory presented in this paper.

Definition 4.1 ("Global" Markov property of DAGs [59])**.** Consider a joint probability distribution, P over a set of variables **V** satisfies the **Markov property** with respect to a graph $G = (V \cup L, E)$ if the following holds for, $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ disjoint subsets of V:

$$P(y|x, z) = P(y|z) \quad \text{if } Y \perp X|Z \text{ in G (that is Y is d-separated from X given Z)}$$

□

The global Markov property maps graphical structure in causal directed acyclic graphs (DAGs) to conditional independence (CI) statements in the relevant probability distributions from data.

Definition 4.2 (Maximal Ancestral Graphs (MAGs) [60])**.** A mixed-edge graph is a maximal ancestral graph (MAG) if:

1. there is no directed cycles
2. there are no almost directed cycles (ancestrality) and
3. there is no primitive inducing path between any two non-adjacent vertices (maximal)

□

Many DAGs may encode the same CI statements, and a MAG encodes an equivalence class of these CI statements that has desirable properties such as maximality and ancestrality. To compare different MAGs, one can leverage Definition 4.3.

Definition 4.3 (General Markov Equivalence from [61])**.** Two MAGs $G_1 = (V, E_1)$, $G_2 = (V, E_2)$ are Markov equivalent if for any three disjoint sets of vertices, X, Y, Z, X and Y are m-separated by Z in G_1 if and only if X and Y are m-separated by Z in G_2. □

Checking Definition 4.3 is quite tedious because it involves explicitly comparing every single m-separation statement possible in both graphs. An equivalent completely graphical criterion in Proposition 2 can be instead used.

Proposition 2 (Graphical Criterion for Markov Equivalence from [61])**.** Two MAGs over the same set of vertices are Markov equivalent if and only if

1. They have the same adjacencies
2. They have the same unshielded colliders
3. If a path p is a discriminating path for a vertex Y in both graphs, then Y is a collider on the path in one graph if and only if it is a collider on the path in the other.

□

Unfortunately, a MAG is not uniquely identifiable (i.e. learnable) from observational data in general. Therefore, a partial ancestral graph (PAG) is defined as the object of interest instead.

Definition 4.4 (Partial Ancestral Graph [60])**.** Let [M] be the MEC of an arbitrary MAG M. The PAG for [M], $\mathcal{P}_{[M]}$ is a partial mixed graph such that:

1. $\mathcal{P}_{[M]}$ has the same adjacencies as M (and any member of [M]) does
2. A mark of arrowhead is in $\mathcal{P}_{[M]}$ if and only if it is shared by all MAGs in [M]
3. A mark of tail is in $\mathcal{P}_{[M]}$ if and only if it is shared by all MAGs in [M].

□

We note that do-calculus is complete for learning PAGs [39]. As noted in [6], the FCI algorithm really only leverages the inversion of R1 within a single domain. If we have access to interventional distributions, the inversion of R2 and R3 enable one to further characterize and learn a more detailed EC [6].

A final lemma due to [62] is useful for proving properties about distributions that satisfy certain graph constraints, but not others. Meek uses the following result to show that set of unfaithful distributions has Lebesgue measure zero.

Lemma 2 (Meek [62])**.** Let $D = (V, E)$ be a causal DAG where $(A \not\perp B|C)_D$. Let $D_s = (V_s, E_s)$ be the subgraph that contains all the nodes in the m-connecting path that induces $(A \not\perp B|C)_D$. Then any distribution p over V_s where every adjacent pair of variables are dependent satisfies $(A \not\perp B|C)_p$. □

As mentioned in 2, compared to a traditional selection diagram, we construct a selection diagram across a set of domains slightly differently. It is a "joint" selection diagram that represents jointly a set of domains and their corresponding S-nodes.

Definition 4.5 (Joint selection diagram)**.** Given a set of domains $\mathbf{\Pi} = \{\Pi^1, ..., \Pi^N\}$ with shared causal structure. For each possible pair of domains, (Π^i, Π^j) there is a selection diagram the contains S-nodes $\mathbf{S}^{i,j}$ that cause changes in the underlying mechanisms. A joint selection diagram is $G = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$, where $\mathbf{S}^{\Pi} = \bigcup_{i,j \in [N], i \neq j} \mathbf{S}^{i,j}$ and $\mathbf{E_S} = \bigcup_{i,j} \mathbf{E}_{\mathbf{S}^{\mathbf{i,j}}}$ is the union of all S-nodes and their edges from each pair of domain's selection diagram. □

D.1.2 Multi-Domain Causal Bayesian Network Invariances

[63] developed an extension of Pearl's do-calculus rules to soft interventions in SCMs. In [64], it was shown that for the general problem of transportability, the generalized do-calculus rules are complete. In this section, we take the do-calculus rules and extend them to invariances present in a Causal Bayesian Network (CBN) that can apply across two arbitrary interventions and two arbitrary domains. This is essential for motivating the S-Markov property characterization and the corresponding equivalence class. This result leads to the Definition 2.2 presented in Section 2.

Lemma 3 (Generalized CBN Invariances Across Domains)**.** Let G be a causal diagram and $G_S = (\mathbf{V} \cup \mathbf{L} \cup S, \mathbf{E} \cup \mathbf{E_S})$ be the corresponding causal selection diagram with latents and S-nodes defined between two domains $\mathbf{\Pi} = \{\Pi^1, \Pi^*\}$ of a CBN. Let $\mathbf{P^{\Pi}}$ be a tuple of interventional distributions generated by G. Let $\mathbf{V_S}$ be the set of nodes that have an edge with respect to $\mathbf{S}$. Then the following distributional invariances hold for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$

(a) For $P^i_{\mathbf{I}} \in \mathbf{P}^{\Pi}$, we have $P^i_{\mathbf{I}}(y|w, z) = P^i_{\mathbf{I}}(y|w)$ if $\mathbf{Y} \perp \mathbf{Z}|\mathbf{W}$ in G.

(b) For $P^i_{\mathbf{I}}, P^j_{\mathbf{J}} \in \mathbf{P}$, we have $P^i_{\mathbf{I}}(y|w) = P^j_{\mathbf{J}}(y|w)$ if $\mathbf{Y} \perp \mathbf{K}|\mathbf{W} \backslash \mathbf{W_K}$ in $G_{\underline{W_K}, \overline{R(W)}}$, where $\mathbf{K} = (\mathbf{I} \Delta \mathbf{J}) \cup \mathbf{V_S^{i,j}}, \mathbf{W_K} = \mathbf{W} \cap \mathbf{K}, \mathbf{R} = \mathbf{K} \backslash \mathbf{W_K}$ and $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ are non-ancestors of $\mathbf{W}$ in G.

Proof. Whenever $i = j$, for domain indicators i, j, then there is no S-node by definition, since the S-node is added to select between different domains i and j. Therefore in constraint (a), because of the shared causal structure assumption 2, $P^i_{\mathbf{I}}(\mathbf{V})$ can just be written as $P_{\mathbf{I}}(\mathbf{V})$ and factorized according to the following equation:

$$P_X(\mathbf{v}) = \sum_{\mathbf{L}} \prod_{i|X_i \in \mathbf{X}} P^*(x_i|\mathbf{pa_i}) \prod_{j|T_j \notin \mathbf{X}} P(t_j|\mathbf{pa_j})$$

which is also known as the truncated factorization formula [1], or the g-formula [65]. Then applying d-separation criterion, constraint (a) follows [66].

Constraint (b) is proven in [67] Thm 4, when $i = j$. So we prove the case when $i \neq j$. To prove this, we take a similar strategy to the proof of the do-calculus rules [1]. We construct a hypothetical CBN that models the selection of a domain on each variable with an endogenous root node/variable along with the intervention on each variable. We assume the change in domain is not caused by any variable in G. Moreover, we assume that soft interventions are triggered by exogenous variables and not affected by any variable in G.

Let $\mathbf{I}^i, \mathbf{J}^j$ denote set of nodes in $\mathbf{I}$ and $\mathbf{J}$ that occur in domains i and j respectively. We can augment G with $\mathcal{F}^{i,j} = \{F_k^{i,j} | V_i \in \mathbf{I}^i \cup \mathbf{J}^j\}$ and edges $\mathcal{E}^{i,j} = \{F_k^{i,j} \to V_k | F_k \in \mathcal{F}\}$.

G_S has an S-node $S^{i,j}$, representing the selection between domain i and j. The edges from $S^{i,j}$ are in $\mathbf{E_S}$ and their direct children are $\mathbf{V_{S^{i,j}}}$. Thus, we constraint b) holds by definition of the selection diagram, if we can remove the effect of the S-node, $S^{i,j}$.

The constructed augmented causal graph is G'. Let Pa_i denote the parents of variable $V_i \in \mathbf{V}$ that excludes nodes in S. Let Pa_i' denote the parents of variables $V_i \in \mathbf{V}$ that can include nodes in S. For each variable V_k with $F_k^{i,j}$, there are a new set of parents $Pa_i'' = Pa_i' \cup \{F_k^{i,j}\}$. The distribution of $P(V_i|Pa_i'')$ is given as follows where $P^l(V_i|Pa_i)$ is a unique conditional probability for each identifier l. We have:

$$P(V_k|Pa_k'') = \begin{cases} P(V_k|Pa_k'), & \text{if } F_k^{i,j} = 0 \\ P^l(V_k|Pa_k'), & \text{if } F_k^{i,j} = l. \end{cases} \tag{4}$$

Furthermore, we can decompose each of those conditional probabilities into ones that are a function of just the Pa_k.

$$P(V_k|Pa_k') = P(V_k|Pa_k), \quad \text{if } S^{i,j} \to V_k \notin E_S \tag{5}$$

and

$$P^l(V_k|Pa_k') = P^l(V_k|Pa_k), \quad \text{if } S^{i,j} \to V_k \notin E_S \tag{6}$$

Thus, each $F_k^{i,j}$ has an arbitrary prior distribution over its domain, which induces a new distribution P'' over $\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \mathcal{F}$ and P'' factorizes according to G'. Then $P'_{\mathbf{I}^i}(\mathbf{V})$ relates to P'' as follows where we condition on every $F_k^{i,j} \in \mathcal{F}$ such that 1) $F_k^{i,j} = 0$ if $V_k \notin \mathbf{I}^i$ and 2) $F_k^{i,j} = l$ if $V_k^l \in \mathbf{I}$.

$$P'_{\mathbf{I}}(\mathbf{V}) = \sum_{\mathbf{L}} P''(\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} | F_k^{i,j} = l, ...)$$

We can similarly decompose P' and relate it to P following the same logic for the S-node. In this sense, we see that the S-nodes and the F-nodes play a similar graphical role in selecting the distribution that applies based on the selection of the S-nodes and F-nodes.

We can repeat the logic for $P_{\mathbf{J}}(\mathbf{V})$. Now, let $\mathbf{F}_{\mathbf{K}}^{i,j} = \{F_k^{i,j} | V_k \in \mathbf{I}^i \Delta \mathbf{J}^j\}$. If $(\{\mathbf{F_K}^{i,j}, S^{i,j}\} \perp \mathbf{Y} | W)_{G'}$, then changing the conditioning values of $\mathbf{F_K}$ and $S^{i,j}$ is irrelevant to $\mathbf{Y}$ and we get $P_{\mathbf{I}}^i(y|w) = P_{\mathbf{J}}^j(y|w)$. Thus we have successfully factorized the two distributions to show they are equivalent when the corresponding graphical criterion holds. □

The result implies that d-separation from S-nodes and their corresponding direct children represent invariances in the conditional probability distributions of observational data assuming the Markov property. Since all S-nodes are source nodes, then to be d-separated from $\mathbf{V_S}$ is equivalent to d-separation from the S-nodes $\mathbf{S}$.

D.1.3 S-Markov Property Results

In this section, we prove some of the results in the main paper from Section 2. The first result proves the statement in Lemma 1.

D.1.3.1 Proof of Main Text Lemma 1 [S-Markov property generalizes the Ψ-Markov property]

Lemma (S-Markov property generalizes the Ψ-Markov property)**.** Let $\mathbf{\Pi} = \{\Pi^1\}$ and $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$. Say $\mathbf{\Psi}^{\Pi}$ and $\mathbf{P}^{\Pi}$ be an arbitrary set of interventions and distributions with $\mathcal{K} = []$. Given $\mathcal{K}$, $\mathbf{P}^{\Pi}$ satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$, then $\mathbf{P}^{\Pi}$ also satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi} \rangle$.

Proof. By assumption, we have only distributions with unknown intervention targets. Given $\mathcal{K}$, $\mathbf{P}^{\Pi}$ satisfies the S-Markov property, so we will show that it also simultaneously satisfies the Ψ-Markov property. Moreover, since there is only one domain $\mathbf{V_S} = \phi$, the empty set.

$$\text{For } \mathbf{I}_i^j \in \mathbf{\Psi}^{\Pi}: \quad P_i^j(y|w,z) = P_i^j(y|w) \quad \text{if } Y \perp Z|W \text{ in D}$$

is satisfied by the first condition of the S-Markov property in Def. 2.2.

$$\text{For } \mathbf{I}_i, \mathbf{I}_j \in \mathbf{\Psi}^{\Pi}: \quad P_i(y|w,z) = P_j(y|w) \quad \text{if } Y \perp Z|W \backslash W_K \text{ in } G_{\underline{W_K}, \overline{R(W)}}$$

is satisfied by the second condition of the S-Markov property. There is only a single domain, so there is by definition no S-node, and thus the condition reduces to the Ψ-Markov property condition two. Therefore, $\mathbf{P}^{\Pi}$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi} \rangle$. □

Lemma 1 demonstrates that the S-Markov property generalizes the Ψ-Markov property. Since the Ψ-Markov property itself has been shown to generalize the I-Markov and Global Markov property, we have the following corollaries.

Corollary 1 (S-Markov property generalizes the I-Markov property)**.** Let $\mathbf{\Pi} = \{\Pi^1\}$, $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, $\mathbf{\Psi}^{\Pi}$ be an arbitrary set of inteventions and $\mathbf{P}^{\Pi}$ an arbitrary set of distributions induced by $\mathcal{I}^{\Pi}$. Let $\mathcal{K}$ be a vector of 1's, such that all distributions have a known intervention target. If $\mathbf{P}^{\Pi}$ satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$, then it also satisfies the I-Markov property with respect to G. □

Corollary 2 (S-Markov property generalizes the Markov property)**.** Let $\mathbf{\Pi} = \{\Pi^1\}$, $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$, $\mathbf{\Psi}^{\Pi} = \langle \{\}^1 \rangle$ and $\mathbf{P}^{\Pi}$ an arbitrary set of distributions. If $\mathbf{P}^{\Pi}$ satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S} \rangle$, then it also satisfies the Markov property with respect to G. □

We defined a joint selection diagram in Definition 4.5. Here, we show that there is no information loss when we construct the joint selection diagram, which is easier to analyze. The joint selection diagram (as defined in Definition 4.5) is a valid representation of a collection of selection diagrams stemming from different domains. Thus we refer to joint selection diagrams as selection diagrams in the main paper.

Lemma 4 (Joint selection diagrams are valid representations)**.** A joint selection diagram preserves transportability phenomena. That is, if a causal effect is transportable in the non-joint selection diagram if and only if it is transportable in the joint selection diagram.

Proof. Since S-nodes are defined as pointing out of S-nodes by construction, then in the joint selection diagram they can act as "confounders" when viewed graphically. Define A as the node that an S-node points to originally. An S-node by definition only has additional edges if there is an inducing path between the A and another node B. If such a path exists, then there is an unblockable subpath from A to B and conditioning on the S-node would not change the m-separation statements. □

In Definition 2.2, we define the Markov property for a (joint) selection diagram. The S-Markov property generalizes the Markov property and extends the conditions of d-separation (condition i) and distributional invariances (condition ii) to selection diagrams. Condition ii is no longer a conditional independence statement, but rather a different type of invariance [68]. Note that compared to the Ψ-Markov property, there are some subtle differences. Namely, there is always the question of whether or not nodes are d-separated with respect an S-node. D-separation with respect to an S-node representing a pair of domains allows one to map invariances across those two domains.

To prove Thm. 1, we first prove a few useful lemmas. The first lemma relates m-separation statements with a conditioning set of S-nodes to other m-separation statements that contain "more" S-nodes.

Lemma 5 (M-separation statements can arbitrarily add S-node singletons)**.** Let G be the joint selection diagram with respect to a causal Bayesian network with latents, $G = (V \cup \mathbf{S}, E \sup E_S)$. Consider m-separation statement with respect to G with $X \perp Y|Z, S_i$ where $X, Y \subset V \cup \mathbf{S}$ and $Z \subseteq V - \{X, Y\}$ and $S_i \subset \mathbf{S} - \{X, Y\}$ (that is S_i is a set of S-nodes).

For any $S_i \in \mathbf{S} - (S \cup \{X, Y\})$, the following statements are equivalent:

1. $X \perp Y | Z, S_i$ in G
2. $X \perp Y | Z, S_i \cup \{S_j\}$ in G and ($S_j \perp Y | Z, S_i$ or $S_j \perp X | Z, S_i$)

Proof. The first statement states that X and Y are d-separated given Z and the ith set of S-nodes in the joint selection diagram.

The second statement states that if we augment the m-separation statement with a conditioning set of the jth S-node, then either the jth S-node is m-separated from Y given Z and S_i or the jth S-node is m-separated from X given Z and S_i.

We show the equivalence of the m-separation statements by analyzing the paths that are m-connecting.

We are given that $X, Y \neq S_j$ and $S_j \notin S_i$. Suppose that there is a m-connecting path between X and Y given Z and S_i (the converse of the first statement in the lemma). Either it passes through S_j S-node or it does not.

If it does not pass through S_j, then since all S_j are oriented out of S_j, then $X \not\perp Y | Z, S_i \cup \{S_j\}$ in G.

If it does pass through S_j, then there are two m-connecting paths that lead from X to S_j given Z and S_i and from S_j to Y given Z and S_i.

If there are no m-connecting paths between X and Y given Z and S_i, then all the paths have to be m-separating. □

Next, we show that when there is a difference in m-separation statements between two selection diagrams, these can be mapped to m-separation statements from U, O, or T, sets that are defined as follows:

We define the following sets of m-separation statements:

$$U = \{(X \perp Y | Z, S)_G \ : X, Y \in V \cup \mathbf{S}, Z \subseteq V - \{X, Y\},\ S \subset \mathbf{S} - \{X, Y\}$$
$$O = \{(X \perp Y | Z, S)_G \ : X, Y \in V \cup \mathbf{S}, Z \subseteq V - \{X, Y\},\ S = \mathbf{S} - \{X, Y\}$$
$$T = \{(X \perp Y | Z, S)_G \ : X \in V, Y \in V \cup \mathbf{S}, Z \subseteq V - \{X, Y\},\ S = \mathbf{S} - \{X, Y\}$$

Intuitively, U, O and T are m-separation statement sets that contain all possible sets of m-separation statements inside a MAG.

Lemma 6 (Arbitrary differences in m-separation statements induce a difference in U, O, or T)**.** Let $G_1 = (V \cup \mathcal{S}, E_1 \cup E_S\}$ and $G_2 = (V \cup \mathcal{S}, E_2 \cup E_S\}$ be selection diagrams over the same sets of variables V. Suppose X, Y, Z are disjoint subsets of $V \cup \mathcal{S}$.

$X \perp Y | Z$ in G_1, $X \not\perp Y | Z$ in G_2, then at least one of the following is true:

i) there exists $X, Y, Z \subseteq V$ such that $X \perp Y | Z, \mathcal{S}$ in G_1 and $X \not\perp Y | Z, \mathcal{S}$

ii) There exists $A, B \subseteq V$ and $S_i \in \mathcal{S}$ such that $(S_i \perp A | B, \mathcal{S} \backslash S_i)$ in G_1 and $(S_i \not\perp A | B, \mathcal{S} \backslash S_i)$ in G_2

In other words: Any difference in m-separation statement from the set of statements $U \cup O \cup T$ between G_1 and G_2 can be stated as just a difference between m-separation statements in T between G_1 and G_2.

Proof. Given m-separation statement in U, we can write these as m-separation statements in O. This is done by repeatedly applying Lemma 5 to m-separation statements in U until all m-separation statements lie in O.

Now, we prove that all m-separation statements in O that are not in T can be mapped to T. First, note that T is a subset of O, since there is the additional constraint that $X \in V$, rather than $X \in V \cup \mathbf{S}$.

Define $W = T \backslash O = \{(X \perp Y | Z, S)_G \ : \ X \in \mathbf{S}, Y \in \mathbf{S}, Z \subseteq V - \{X, Y\}$ as the set of m-separation statements that are in O, but not in T. These are m-separation statements then between S-nodes of the selection diagram. We consider any m-separation statement where $S_i \perp S_j | Z, \mathbf{S} - \{S_i, S_j\}$ in G_1, but $S_i \not\perp S_j | Z, \mathbf{S} - \{S_i, S_j\}$ in G_2.

S-nodes are by Definition 4.9 pointing out, there must be at least one collider along paths between S_i, S_j. First we consider a path that is active in G_2, but not in G_1. If $S_i \perp S_j | Z, \mathbf{S} - \{S_i, S_j\}$ in G_1, but $S_i \not\perp S_j | Z, \mathbf{S} - \{S_i, S_j\}$ in G_2 for some $Z \subset V$, then this can only happen if in G_2, there exists a node in Z that is a descendent of both S_i and S_j. That is $v \in Z$ such that $v \in Desc(S_i) \cap Desc(S_j)$, which makes the collider active in G_2. In G_1, we have simultaneously that for all nodes in Z, there does not exist any descendants of both S_i and S_j. That is $\nexists v \in Z$ such that $v \in Desc(S_i) \cap Desc(S_j)$. This then means that v is either not a descendant of S_i, or it is not a descendant of S_j. Suppose WLOG that v is not a descendant of S_i.

Then this implies that $S_i \perp v | \mathbf{S} - \{S_i\}$ in G_1, and $S_i \not\perp v | \mathbf{S} - \{S_i\}$ in G_2.

Now, suppose $(X \perp Y | Z)_{D_1}$ and $(X \not\perp Y | Z)_{D_2}$. Any m-separation statment belongs to one of the sets O, U or T. Since G_1 and G_2 share the same vertex set, V, then the m-separation statement would be in the same set.

If this m-separation statement set belongs to T, then we are done.

If it belongs to O, then by our earlier result, any m-separation statement with differences imply an m-separation statement difference in T and the result follows.

If it belongs to U, then by Lemma 5 and the above, the m-separation statement can be mapped to m-separation statements in O. Then by the previous statement, the result follows.

This proves the lemma. □

D.1.4 Results from Section 2.3 on observational multi-environment Markov equivalence

In the following results leading up to Theorem 1, we assume that we only have access to observational data across multiple domains. In this setting, the S-Markov property and the relevant graphical S-Markov equivalence properties are much simpler. We show that there is a mapping at this point between the S-Markov property and the Ψ-Markov property [6]. We are ready to prove an equivalent graphical condition for S-Markov equivalence.

Theorem 4 (Graphical S-Markov Equivalence Among Selection Diagrams With Only Observational Data)**.** Let G^1 and G^2 be two causal diagrams. Let $G^1_S = (V \cup L_1 \cup S_1, E_1)$ and $G^2_S = (V \cup L_2 \cup S_2, E_2)$ be their corresponding selection diagrams over N environments with S-nodes $\mathbf{S_1}, \mathbf{S_2}$, $\mathbf{\Pi} = \{\Pi^1, ..., \Pi^N\}$, with interventions $\mathbf{\Psi}^{\Pi} = \langle \{\}^1, \{\}^2, ..., \{\}^N \rangle$ and associated distributions $\mathbf{P}^{\Pi}$. Let $\mathcal{K} = [1, 1, ..., 1]$ be the vector of known interventions.

We say $\langle G^1_S, \mathbf{\Psi}, \mathbf{S} \rangle$ and $\langle G^2_S, \mathbf{\Psi}, \mathbf{S} \rangle$ are S-Markov equivalent if and only if for $M_1 = MAG(Aug_{\Psi, V_S}(G^1)$ and $M_2 = MAG(Aug_{\Psi, V_S}(G^2)$:

1. M_1 and M_2 have the same skeleton;
2. M_1 and M_2 have the same unshielded colliders;
3. If a path p is a discriminating path for a node Y in both M_1 and M_2 then Y is a collider on the path in one graph if and only if it is a collider on the path in the other.

Proof. (=>) Assuming that $MAG(D_1)$ and $MAG(D_2)$ satisfy the three conditions, we will show they are S-Markov equivalent. Then by Definition 4.3 and Proposition 2, the two MAGs have the same m-separation statements and vice versa, thereby satisfying the S-Markov equivalence condition, where both G^1_S and G^2_S impose the same constraints over the set of distributions defined in 4.3.

(<=) We prove this direction by contradiction. Suppose $MAG(D_1)$ and $MAG(D_2)$ do not satisfy the three conditions, then we want to show that the two graphs are not S-Markov equivalent.

By definition of a MAG, if the two MAGs have one of the conditions different, then there is at least one different m-separation statement. Without loss of generality, we consider only m-separation statements among pairs of singletons. If an m-separation statement holds between arbitrary sets of nodes in one selection diagram, G_1, but not G_2, then there is at least one pair of singletons where the m-separation statement differs between G_1 and G_2.

Consider the sets U, O and T again of m-separation statements in Lemma 6. U, and O, are m-separation statements between any two nodes given a strict subset of all remaining S-nodes, all

remaining S-nodes. T is the set of m-separation statements between normal nodes and any other node given all remaining S-nodes.

By Definition 2.2, an m-separation statement is in T if and only if it appears in the S-Markov equivalence class of distributions for G.

By Lemma 6, we show that if the two MAGs of the selection diagram are not Markov equivalent, then there is a m-separation statement in the definition of S-Markov equivalence that is different in the two graphs. As a result, we are able to show that there is a tuple $\langle G^2_S, \mathbf{\Psi}, \mathbf{S} \rangle$ that contains tuples of distributions $\mathbf{P}$ that is not S-Markov with respect to $\langle G^2_S, \mathbf{\Psi}, \mathbf{S} \rangle$. □

Thus, graphically, the two selection diagrams over multiple domains of only observational data are Markov-equivalent if the MAGs of their augmented diagrams fulfill certain similarity constraints.

D.1.4.1 Proof of Main Text Thm. 1 [Equivalence of Ψ and S-Markov property given multi-domain observational distributions]

Next, we state an equivalence between the Ψ-Markov characterization in Theorem 1 of [6] and S-Markov characterization.

Theorem (Equivalence of Ψ and S-Markov property given multi-domain observational distributions)**.** Let G be a causal diagram and $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ the selection diagram over N domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ...\Pi^N\}$. Let $\mathbf{S}^{\Pi}$ be set of S-nodes and $\mathbf{E_S}$ their set of edges. Let $\mathbf{\Psi}^{\Pi} = \langle \{\{\}\}^1, ..., \{\{\}\}^N \rangle$ and $\mathcal{K} = [1, 1, ..., 1]$, such that for each of the N domains, there is only observational data. Let $\mathbf{P}^{\Pi}$ be an arbitrary set of distributions. If $\mathbf{P}^{\Pi}$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S} \rangle$, then it also satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S} \rangle$.

Proof. If $\mathbf{P}$ satisfies the Ψ-Markov property, then for disjoint $Y, Z, W \subseteq V$ the condition related to d-separation of is held for each distribution in the joint selection diagram given the shared causal structure assumption.

For the second condition relating pairs of distributions to each other in the Ψ-Markov property, we know that this is equivalent to d-separation in the augmented graph with the augmented graph nodes added from pairs of different distributions given the Definition 2.3. In our case, each pair of distibutions correspond to a pair of different domains, and thus the augmented F-node has a similar meaning to the S-node.

Let Z be a S-node (that is represented by an F-node in the augmented graph) and $Y \perp Z | W, S_{[N]\setminus[i]}$. This then shows that if the Ψ-Markov property holds, then the S-Markov property holds. Similarly if the S-Markov property holds with respect to $\langle G, \mathbf{\Psi}, \mathbf{S} \rangle$, then it implies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi} \rangle$. □

This proves the result stated in Thm. 1 and shows that the Ψ-Markov property implies the S-Markov property in the case where only observational data is present in multiple domains. This can be seen conceptually that the domain change can be viewed as an intervention on the data distributions with unknown targets (i.e. we do not know where the environment targets). In fact they are equivalent in this setting.

Corollary 3 (An equivalence of S-Markov Equivalence and Ψ-Markov Equivalence)**.** Let G be a causal diagram and $G_S = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ the selection diagram over N domains $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ...\Pi^N\}$. Let $\mathbf{S}^{\Pi}$ be set of S-nodes and $\mathbf{E_S}$ their set of edges. Let $\mathbf{\Psi}^{\Pi} = \langle \{\{\}\}^1, ..., \{\{\}\}^N \rangle$ and $\mathcal{K} = [1, 1, ..., 1]$, such that for each of the N domains, there is only observational data. Let $\mathbf{P}^{\Pi}$ be an arbitrary set of distributions. $\mathbf{P}^{\Pi}$ satisfies the Ψ-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S} \rangle$, if and only if it satisfies the S-Markov property with respect to $\langle G, \mathbf{\Psi}, \mathbf{S} \rangle$.

Proof. The result follows from 1 and D.1.4.1. □

This proves an interesting equivalence mapping between multi-domain observational setting and single-domain unknown interventional setting. One can view the change in domain as an unknown intervention that occurs via nature. However, knowing the domain change is still import information as not only does nature induce an intervention, but there may also be various interventional datasets collected explicitly in the domain. Thus one would know that these interventions in this domain are different from similar interventions in another domain. Note there are a few subtle differences that one should be aware of.

1. In the case of Ψ-Markov equivalence, one works with an augmented graph with the symmetric difference between all pairs of different intervention target sets. The sets of variables from the symmetric difference of intervention targets form the "F-nodes" of the augmented graph, which can then be viewed analagously to S-nodes. In S-Markov equivalence, one has S-nodes on all variables that have differences between the source and target domains. These S-nodes represent possible distribution differences between pairs of domains regardless of whether or not the distributions are observational or interventional. Thus S-nodes can be seen as "nature's intervention" that is always present.
2. S-nodes represent a difference in distribution between the source and target domain at the nodes it points to. An F-node represents a difference between a pair of distributions due to a symmetric difference in intervention target. These are important subtleties, which allow the user to utilize such qualitative information in downstream transportability ID tasks [17]. The extra information that comes from the knowledge that each observational distribution comes from a different environment manifests purely in the interpretation of the nodes. However, transportability ID in an EC is still an open problem, and thus it is unclear how to leverage the results of the learning algorithm.

Based on this equivalence of S-Markov and Ψ-Markov property for multi-domain observational data, the Algorithm S-FCI introduced in this paper is sound and complete. That is every adjacency and orientation is common for all $MAG(G')$ where G' is a selection diagram S-Markov equivalent to G. Moreover the recovered graph is the most informative it can be (i.e. discovers as many tails and arrowheads that can be oriented within a S-Markov equivalence class).

Corollary 4 (Modified Ψ-FCI algorithm to learn an S-PAG)**.** Define the modified Ψ-FCI algorithm with two modifications: i) represent S as the set of intervention targets and ii) take the graph learned and remove all S-nodes that represent a pairing between distributions from two target domains. The modified Ψ-FCI algorithm is complete for learning an S-PAG given only observational data.

Proof. If we run Ψ-FCI, with the S-nodes represented as our intervention targets, then we will learn a supergraph of the graph of interest. The supergraph will contain extra F-nodes due to symmetric differences among the combinations of source domains. By removing those, we have a Ψ-PAG with only F-nodes representing the source and target domain, which is the S-PAG. □

D.1.5 Results from Section 2.4 obs. + interv. data in multiple domains

In this section, we prove results related to causal discovery in the setting of multiple domains with observational and interventional data. First, we show some equivalence relations when going from the non-augmented graph to the augmented graph.

Proposition 3 (augmented graph Equivalence Relations)**.** Let $G = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S}, \mathbf{E} \cup \mathbf{E_S})$ be a joint selection diagram, with latent variables L and its augmented graph $Aug_{\mathbf{\Psi},\mathbf{S}}(G) = (\mathbf{V} \cup \mathbf{L} \cup \mathbf{S} \cup \boldsymbol{\mathcal{F}}, \mathbf{E} \cup \mathbf{E_S} \cup \boldsymbol{\mathcal{E}})$ with respect to the intervention set across all N domains $\mathbf{\Pi}$, where $\mathcal{F} = \{F_i^j\}_{i\in[k]}^{j\in[N]}$. Let A_{ij} be the set of nodes adjacent to F_i^j for all $i \in [k]$ and all $j \in [N]$. And denote B_i as the set of nodes adjacent to $S^{i,j} \in \mathbf{S}$. We have the following equivalence relations:

For disjoint $Y, Z, W \subseteq V$, we have:

$$(Y \perp Z|W)_G \iff (Y \perp Z|W, F_{[k],[N]})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{7}$$

For each A_{ij} suppose $Y, W \subseteq V \backslash G_{ij}$, we have:

$$(Y \perp A_{ij}|W)_{G_{\underline{A_{ij}}}} \iff (Y \perp F_{ij}|W, A_{ij}, F_{[k]\backslash\{i\}}^{[N]\backslash\{j\}})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{8}$$

$$(Y \perp A_{ij}|W)_{G_{\overline{A_{ij}(W)}}} \iff (Y \perp F_{ij}|W, F_{[k],[N]})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{9}$$

For each A_{ij}, let $Y, W \subseteq V$, and let $W_{ij} = W \cap A_{ij}$, $R = A_{ij} \backslash W_{ij}$, then:

$$(Y \perp A_{ij}|W \backslash W_{ij})_{G_{\underline{A_{ij}}}} \iff (Y \perp F_{ij}|W, A_{ij}, F_{[k]\backslash\{i\},[N]\backslash\{k\}})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{10}$$

For each $S_i \in S$, let $Y, W \subseteq V$, then

$$(Y \perp S^{i,j}|W)_G \implies (Y \perp B_{ij}|W, S^{[N]\times[N]})_{Aug_{\mathbf{\Psi},\mathbf{S}}(G)} \tag{11}$$

Proof. Conditioning on a source node is equivalent to removing it from the graph in terms of the graph separation statements. Hence, conditioning on $F_{[k]\setminus i,[N]\setminus\{j\}}$ in the right-hand side eliminates them. Therefore, equations 7, 8, 9 and 11 follow from [[69], Proof of Th. 4.1] by Pearl.

Note that $S_j \perp F_{ij}$ for all $i \in [k], j \in [N]$ because S-nodes and F-nodes in this setting are source nodes and thus will always have a collider due to the multi-domain intervention assumption.

The rest of the proof is exactly as it is in [Proposition 3 of [7]]. □

The proof for Prop. 1 follows.

D.1.5.1 Proof of Main Text Proposition 1 [Graphical S-Markov Property]

Proposition 4 (Graphical S-Markov Property)**.** Consider the multi-domain setup 1.1. Let $M = MAG(Aug_{\Psi,V_S}(G))$ and let $[M]$ be the set of S-MAGs corresponding to all the triplets $\langle G', \mathbf{\Psi'}^{\Pi}, \mathbf{V_S}\rangle$ that are S-Markov equivalent to $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S}\rangle$. The S-PAG for $\langle G, \mathbf{\Psi}^{\Pi}, \mathbf{V_S}\rangle$, denoted $\mathcal{P}$ is a graph such that:

$$(Y \perp Z|W)_{G_S} \iff (Y \perp Z|W, F_{\mathcal{E}}, \mathbf{S})_{Aug_{\Psi,V_S}(G)} \tag{12}$$

$$(Y \perp \{S^{j,k}, \mathbf{K_i^{j,k}}\}|W\setminus W_i^{j,k})_{G_{S\underline{W_i^{j,k}},\overline{R(W)}}} \iff$$
$$(Y \perp \{S^{j,k}, F_i^{j,k}\}|W, \mathbf{S}\setminus S^{j,k}, F_{\mathcal{E}}\setminus F_i^{j,k})_{Aug_{\Psi,V_S}(G)} \tag{13}$$

where $\mathbf{W}_i^{j,k} = \mathbf{W} \cap \mathbf{K}_i^{j,k}, \mathbf{R} = \mathbf{K}_i^{j,k}\setminus\mathbf{W}_i^{j,k}$.

Proof. The proof follows from Proposition 3. □

We see that graphical equivalence is nicely modular using the augmented graph framework, where we add nodes indicating change in distributions due to domain, or interventions. Similarly, since the augmented graph is still a DAG and the MAG of the augmented graph is a MAG, and the corresponding PAG of the augmented graph is a PAG, we can leverage existing theory that analyzes properties of those graphs. Next, we prove Thm. 2 showing a graphical criterion for determining the S-Markov equivalence among two graphs.

D.1.5.2 Proof of Main Text Thm. 2 [S-Markov Characterization]

Theorem (S-Markov Characterization)**.** Let there be two causal graphs $G^1 = (\mathbf{V} \cup \mathbf{L}_1, \mathbf{E}_1)$, $G^2 = (\mathbf{V} \cup \mathbf{L}_2, \mathbf{E}_2)$ with G_S^1 and G_S^2 the selection diagrams and a corresponding set of intervention targets, $\mathbf{\Psi_1}^{\Pi}$, $\mathbf{\Psi_2}^{\Pi}$, a corresponding set of S-nodes set $\mathbf{S_1}^{\Pi}, \mathbf{S_2}^{\Pi}$ and a fixed index vector of known intervention targets $\mathcal{K}$. Assume that the symmetrical difference sets are indexed in both sets in the same pattern such that correspondence between F-nodes and S-nodes are the same in M_1 and M_2. Then $\langle G_S^1, \mathbf{\Psi_1}^{\Pi}, \mathbf{S_1}^{\Pi}\rangle$ and $\langle G_S^2, \mathbf{\Psi_2}^{\Pi}, \mathbf{S_2}^{\Pi}\rangle$ are S-Markov equivalent if and only if for $M_1 = MAG(Aug_{\mathbf{\Psi_1},\mathbf{S_1}}(G^1)$ and $M_2 = MAG(Aug_{\mathbf{\Psi_2},\mathbf{S_2}}(G^2)$:

1. M_1 and M_2 have the same skeleton
2. M_1 and M_2 have the same unshielded colliders
3. If a path p is a discriminating path for a node Y in both M_1 and M_2, then Y i s a collider on the path in one graph if and only if it is a collider on the path in the other.

Proof. We proved a similar version earlier in Thm. 4 for only multi-domain observational data.

(<=) Suppose the two MAGs, M_1, M_2 satisfy the three conditions stated. Then, they induce the same m-separation statements [61]. Therefore, by Prop. 1, G_1 and G_2 impose the same constraints over the distributions in the S-Markov property definition (Def. 2.2). Therefore, $S^{\Pi}_{\mathcal{K}}(G_1) = S^{\Pi}_{\mathcal{K}}(G_2)$.

(=>) Suppose by way of contradiction that the two MAGs do not satisfy the three conditions. Then at least one different m-separation statement is present, since the MAGs encode m-separation statements. With a different m-separation statement, we want to show they are also therefore not S-Markov equivalent.

In Lemma 6, we demonstrated that any m-separation statement is included in the defined sets $U \cup O \cup T$. Therefore, there is an m-separating path in one graph that is m-connecting in the other. In the final step, we demonstrate that the distribution tuple of Def. 2.2 is different in $Aug(G_1)$ vs $Aug(G_2)$.

We do this by construction.

Suppose $X, Y, Z \subseteq V$ such that $(X \perp Y | Z, \mathcal{F}, \mathbf{S})_{AugG_1}$ and $(X \not\perp Y | Z, \mathcal{F}, \mathbf{S})_{AugG_2}$. Any tuple of distributions (observational or interventional) across any domain obtained is faithful to the selection diagram with latent variables will suffice to demonstrate the proof.

Suppose $X = F_i^j$ for some $i \in [k]$ and $Y \in V$. Therefore an F-node is m-connected to an observed variable in $Aug(G_2)$ but not in $Aug(G_1)$. Now, consider $G_{path} = (V_{path}, E_{path})$, the subgraph of G_2 that includes all the variables that contribute to the m-connecting path of $(X \not\perp Y | Z, \mathcal{F}, \mathbf{S})_{Aug(G_2)}$.

Consider now a jointly Gaussian distribution, p_{path}, on V_{path} that is faithful to G_{path}. Thm. 7 of [62] shows that this is possible.

We proceed now by considering two interventions I, J on the graph where $I \Delta J = A_i$, where the distributions p_I, p_J from the same domain are responsible for the graphical separation of F_i^j. Different from the rest of the paper, for this proof we will treat F_i^j as a regime variable that indicates when we switch to p_I and when we switch to p_J. Note that we can do this since we only add this single F node and no others in this domain j. Consider the distribution $p*$ defined as follows: $p*(.|F_i^j = 0) = p_I(.), p*(.|F_i^j = 1) = p_J$.

We will now show that the variable F_i^j is dependent with Y given Z on the distribution $p*$. So, we construct a SCM that induces an interventional distribution and the relevant graph in question.

Consider the following linear SCM: $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where A is a lower-triangular matrix that captures the DAG structure and parent-child relationships in G_{path} and $\mathbf{e} \in \mathbb{R}^d$ is an exogenous noise vector and d is the number of observed variables in the graph. Let p_I be the distribution obtained by adding noise vector $\mathbf{e}_I$ to the system. $\mathbf{e}_I$ is non-zero in the rows corresponding to the nodes that it perturbs. Therefore p_I is a valid soft-interventional distribution. Let $\mathbf{e}_J$ be the noise vector now for adding an intervention on J.

Next, we show that every adjacent variable is dependent. The correlation of variables in G_{path} is computed as:

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{e} + \mathbf{e}_I \implies (\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{e}_1 \implies \mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{e}_1$$
$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{e} + \mathbf{e}_J \implies (\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{e}_2 \implies \mathbf{x} = (I - A)^{-1}\mathbf{e}_2$$

with $\mathbf{e}_I = \mathbf{e} + \mathbf{e}_1$ and $\mathbf{e}_J = \mathbf{e} + \mathbf{e}_2$. Note when $\mathbf{e}_1$ and $\mathbf{e}_2$ are different, then the F-variable is dependent with the variables in $K := I \Delta J$, since $p(K|F = 0) \neq p(K|F = 1)$, implying $(K \not\perp F)_{p*}$. We can compute the correlation matrix between observed variables with respect to $p*(.)$, since the binary regime variable can be marginalizd out:

$$E[\mathbf{x}\mathbf{x}^T] = 0.5(\mathbf{I} - \mathbf{A})^{-1} E[\mathbf{e_1}\mathbf{e_1}^T](\mathbf{I} - \mathbf{A})^{-1^T} + 0.5(\mathbf{I} - \mathbf{A})^{-1} E[\mathbf{e_2}\mathbf{e_2}^T](\mathbf{I} - \mathbf{A})^{-1^T} \quad (14)$$
$$= 0.5\mathbf{A})^{-1}(\mathbf{D_1} + \mathbf{D_2})(\mathbf{I} - \mathbf{A})^{-1^T} \quad (15)$$

where $\mathbf{D}_i = E[\mathbf{e_i}\mathbf{e_i}^T]$ are the diagonal covariance matrices of the noise that is added due to the interventions. Now, consider two adjacent variables $x_i, x_j \in V_{path}$. Since x_i, x_j are jointly Gaussian,

they are dependent if and only if they are correlated. Therefore, we want to show $E[x_i x_j] \neq 0$ for any arbitrary adjacent pairs. By assumption, any pair of adjacent variables are dependent since the original distribution is chosen to be faithful to the graph G_{path}. Therefore, if we randomly pick variances of the added noise terms, with probability 1, any adjacent pairs of variables will be dependent still after a union bound.

Therefore, in the graph G_{path} plus the F-variable for interventions specifically in domain j, every pair of adjacent variables are dependent. Then, using Meek's Lemma 2, we have that $(F_i^j \not\perp Y|Z)_p$. Since we have not added any other F-variables in this domain as regime variables, we do not need to condition on them. Now, we can augment this distribution to cover variables outside G_{path}. Pick all remaining variables independent from the variables in G_{path} and construct interventional distributions by adding extra noise terms to the intervened variables. Note, even with different domains, we only need to construct a distribution valid within that specific domain. That is, being adjacent to an F-node not associated with the particular domain in question, is irrelevant.

We can repeat the same procedure as described for S-nodes that are now adding extra noise terms to the nodes in which it changes the distribution due to the domain. We simply pick one of the observational distributions as a reference and fix it. Then we arbitrarily add noise terms to each node that has an S-node that perturbs the distribution with respect to another domain. We continue until all domains have an associated distribution.

All that is left is to account for the case, where we are comparing interventions between different domains. That is, we must account for simultaneously a S-node and F-node change in regime. Without loss of generality, we can consider the case of just two domains, i and j and interventions I from domain i and J from domain j. In this case, we can define $\mathbf{e}_1 = \mathbf{e} + \mathbf{e}_I + \mathbf{e}_S$ and $\mathbf{e}_2 = \mathbf{e} + \mathbf{e}_J + \mathbf{e}_S$, where $\mathbf{e}_S$ is the noise vector added due to the change in domain. It is defined with non-zero values at the rows of nodes that are affected by the S-node $S^{i,j}$. Since $\mathbf{e}_S$ is constant between both $\mathbf{e}_1$ and $\mathbf{e}_2$, we can simply redefine $\mathbf{e}' = \mathbf{e} + \mathbf{e}_S$ and the result still follows. Then to generalize across all possible domains, we fix a domain and observational distribution and repeat the process until all domains have an associated set of observational and interventional distributions.

The corresponding tuple of distributions across interventions and domains belong to $S^{\mathbf{\Pi}}_{\mathcal{K}}(G_1)$, but not $S^{\mathbf{\Pi}}_{\mathcal{K}}(G_2)$ since m-separation should have implied invariance between the interventional and domain-change distributions whereas we constructed the distributions such that this is not true. □

The difference between this statement and the one is Thm. 4 is simply what data is available. But the Lemma 6 does not care what sort of data is available, but is simply a result of the graphical structure.

Given Thm. 1, we can leverage the Ψ-FCI algorithm in the multi-domain observational data setting.

Corollary 5 (Modified Ψ-FCI algorithm given multi-domain observational data)**.** Let $\mathbf{\Pi} = \{\Pi^1, ..., \Pi^N\}$ be N domains with $\mathbf{P}^{\mathbf{\Pi}}$ generated from $\mathbf{\Psi}^{\mathbf{\Pi}} = \langle \{\}^1, \{\}^2, ..., \{\}^N \rangle$ consists of N observational distributions. Define the modified Ψ-FCI algorithm with the following modification: represent S as the set of intervention targets. The resulting Ψ-PAG learned is the same as the S-PAG. □

Therefore, the Ψ-FCI algorithm is applicable to the multi-domain setting when there is only observational data.

In the final theorem, we show that the S-FCI algorithm is sound, in that it learns a valid S-PAG (i.e. PAG with additional orientations).

D.1.5.3 Proof of Main Text Thm. 3 [S-FCI Soundness]

Theorem (S-FCI Soundness)**.** Assuming tuple $\mathbf{P}^{\mathbf{\Pi}}$ is generated by some unknown tuple $\langle G, \mathbf{\Psi}^{\mathbf{\Pi}}, \mathbf{S}^{\mathbf{\Pi}} \rangle$ with known intervention target $\mathcal{K}$ from domains $\mathbf{\Pi}$ and is s-faithful, where $\mathbf{\Psi}$ is a tuple of set of interventions with known/unknown targets, $\mathbf{S}$ and its corresponding edges indicate the S-nodes and their edges and G is the causal diagram, with G_S being the selection diagram. S-FCI algorithm is sound (i.e. every adjacency and orientation in $\mathcal{P}_{S-FCI}$ is common for $MAG(Aug_{\mathbf{\Psi},\mathbf{S}}(G))$).

Proof Idea. In order to prove soundness that the result of S-FCI is a valid S-PAG, we will show that the algorithm's inferred separating sets between pairs of nodes are valid.

We determine:

1. Are all pairs of separable nodes in the graph correctly identified? I.e. all edges in the PAG are the result of an adjacency in the underlying DAG, or a primitive inducing path.
2. Do the augmented separating sets affect (negatively) the application of FCI rules?
3. Are the additional orientation rules sound?

The proof idea for the additional orientation rule R9' is as follows: adjacencies in a MAG are due to either adjacency in the true underlying selection diagram, or an inducing path between two nodes. Determining when this inducing path is the case across multiple domains and different interventions with known-targets allows one to then orient this inducing path. □

Proof. (1) All pairs of F-nodes and S-nodes are separable with the empty set by construction of the augmented graph. Hence, after phase I of the S-FCI algorithm, they are non-adjacent with $SepSet(F_{ij}, F_{kl}) = \phi$ and the same for the S-nodes.

(2) To validate that the existing FCI rules are sound, we simply need to check that the rules that rely on separating sets are still valid given our augmented separating sets. The orientation of unshielded colliders and discriminating paths is sound based on the same reasoning as that in [7], since S-nodes are also in fact source nodes.

(3) Finally, we address the soundness of orientation rules. In [7] R9 of the $\mathcal{I}$-FCI algorithm is proved sound, which we follow a similar logic.

Define A_{ijk} as the set of nodes that are children of the F-nodes $F_i^{j,k}$.

We consider a pair of nodes $F_i^{j,k}, Y$, where $F_i^{j,k} \in \mathcal{F}^{\Pi}, Y \in V$ that are not adjacent, but $Y \in Neigh(A_{ij})$, indicating that there is no separating set between $F_i^{j,k}$ and Y in the augmented graph. Since they are not adjacent by construction, then there must be an inducing path between the two nodes relative to latent variables L. The same argument applies to separate Y from $S^{j,k}$. Therefore the MAG of the augmented diagram, $MAG(Aug_{\mathbf{\Psi},\mathbf{S}}(G))$ contains an edge from this node to Y. □

D.1.6 Results improving efficiency of skeleton discovery phase

The skeleton discovery phase of the $\mathcal{I}$-FCI and Ψ-FCI algorithm require testing every possible combinations of nodes with every possible combination of conditioning sets. Constraint-based causal discovery algorithms typically searches for invariances by testing for example conditional independences among existing node pairs. These algorithms then typically may test all possible nodes as part of the separating set.

In the $\mathcal{I}$-FCI [7] and Ψ-FCI algorithms [6], the algorithms compare run through every single possible conditioning set when comparing distributions similar to the SGS algorithm [2]. However, this is obviously very inefficient.

This strategy while sound and works in theory, is very inefficient. Other strategies involve considering only neighbors, such as in the PC algorithm [5]. In addition, the FCI algorithm has been extended to be more computationally efficient, by only testing the possibly d-separating sets [42]. When dealing with the augmented graphs, we would like to ignore the augmented nodes that are irrelevant in the conditioning set. This is possible because we will see graphically that none of the augmented F-nodes constructed in Def. 2.3 are part of the possibly d-separating sets between nodes.

This enables one to speed up the S-FCI algorithm during the skeleton discovery stage using the same techniques.

Definition 4.6 (Possible-D-sep sets)**.** Let G be a mixed-edge graph with circular endpoints, and bidirected edges. $pds(X, Y)$ in G is defined as follows:

$X \in pds(G, X, Y)$ if and only if there is a path π between X and Y in G such that every subpath (X_i, X_j, X_k) of π, X_j is a collider on the subpath in G, or (X_i, X_j, X_k) is a triangle in G. □

The $pds(X, Y)$ set is useful because $pds(X, Y) \supseteq dsep(X, Y)$.

Lemma 7 (S-nodes are not required to be part of a d-separating sets)**.** Let $G = (V \cup S, E \cup E_S)$ be a joint selection diagram. Define $PDS(X, Y)$ as the possibly d-separating sets of X and Y as defined in [42]. For all $X, Y \subseteq V$ disjoint, no S-nodes are required to be part of d-sep(X, Y).

Proof. Assume an S-node, S_i is only pointing to one node, $Z \in V$. Then it is always an ancestor of Z. Consider disjoint $X, Y \subseteq V \backslash \{Z, S_i\}$. For X and Y to be d-separated, all d-connected paths must be blocked. Consider the path from X to Y through Z. If the path is a collider at $A \star\!\!\rightarrow Z \leftarrow\!\!\star B$, then the triplet (A, Z, B) is blocked as long Z, or descendants of Z are not conditioned on. If the path is a non-collider at Z, then it is blocked as long as Z is conditioned on. In both scenarios, S_i may be added to the conditioning set without changing the blocked/unblocked status of the triplet.

Consider now $S_j \in S$ that is another S-node. If that S-node is not along the path from X to Y, then it can be conditioned on arbitrarily since it is never a descendant of a collider and therefore would not open up a collider path.

Say S_i is pointing to now multiple nodes due to inducing paths. The argument is the same now for each node it is pointing to. If S_i is pointing to multiple nodes, the presence of an inducing path between X and Y indicates that there is no d-separating set between X and Y, so even if S_i is a graphical "confounder", adding it or not would not change the d-connectedness between X and Y. □

This is useful to know as the skeleton search phase of the FCI algorithm and its variants typically rely on defining a superset of the d-separating set between variables, such as the $PDS(X, Y)$ set. Based on this lemma, we do not need to include any S-nodes ever in the conditioning set. This results in a faster skeleton discovery stage in S-FCI, which we incorporate into our implementation.

The skeleton phase proceeds as follows:

1. Run the FCI skeleton discovery phase among the non S-node variables using neighbors to select the conditioning sets
2. Orient unshielded colliders
3. Compute the $pds(X, Y)$ for all disjoint $X, Y \in V$
4. Orient all edges into circular endpoints
5. Re-run the FCI skeleton discovery phase using the $pds(X, Y)$ to select conditioning sets
6. Repeat the above now among S-node variables and non-S node variables

See Algorithm D.3 where we can leverage the strategy of possibly d-separating sets in the "CondSel" function. Moreover, we can limit the PDS set further by always removing all S-nodes and F-nodes from the PDS set.

D.2 Learning Selection Diagrams Across More Than Two Domains

The traditional selection diagram is presented with S-nodes that represent a change in mechanism between a pair of domains [17]. When we extend this to allow more than two domains, we can add additional S-nodes for each pair of domains. Consider Figure S1(a), where there are three S-nodes representing domains 1, 2 and 3. The presence of an S-node edge means there is not necessarily an invariance of X: i.e. $P^i(X) = P^j(X)$ is not necessarily true for $i \neq j$. However, in Figure S1(b), removing the edge between $S^{1,2}$ and X indicates that an invariance is present in the marginal distribution, $P^1(X) = P^2(X)$. However, if we also remove the edge $S^{1,3} \rightarrow X$, then this implies the invariance $P^1(X) = P^3(X)$. Then by transitivity, $P^2(X) = P^3(X)$ must also be true and the S-node edge $S^{2,3} \rightarrow X$ should also be removed in order for the graph to be valid.

This removal means that with higher number of domains, the learning of invariances across domains due to the lack of S-node edges can be accelerated. Say we have observational data across three domains and the selection diagram indicates that X is d-separated from all the S-nodes. Then as soon S-FCI determines the invariance such that the corresponding F-node, $F^{i,j}_{\{\}}$ is removed for two pairs of domains (1,2) and (1,3), then it can immediately remove the F-node $F^{2,3}_{\{\}}$ since the invariance must be true as well. To determine the invariant domains per node in the graph, one simply needs to construct an undirected graph among the domain IDs of the removed S-node edges and compute the connected components, which can be done in $\bigoplus(N)$ time, where N is the number of domains. This is a common graph algorithm that uses a disjoint set and is implemented in a variety of different packages, such as networkx [70]. This enables one to efficiently compute the invariant domains during the skeleton removal phase of Algorithm D.3.

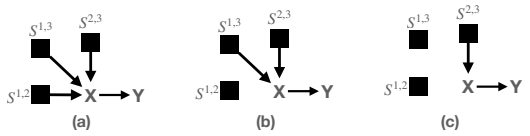

Figure S1: (a) shows a selection diagram with 3 domains with the distribution of X changing across any pair of domains. If we remove an edge $S^{1,2} \to X$, then this implies that for domain 1 and 2, the distribution of X is invariant (b). However, if we also remove the edge $S^{1,3} \to X$, then this additionally implies for domain 1 and 3 the distribution of X is invariant (c). Without explicitly testing the invariance, one can remove the edge $S^{2,3} \to X$ by transitivity. The reasoning is described in more detail in Section D.2.

This improvement due to limiting the necessary CI tests needed to be run can help improve runtime of the S-FCI algorithm.

D.3 Comparing Markov Properties

We compare the different Markov properties in greater detail here. The Markov property maps graphical d-separation to invariances in the decomposition of the joint probability distribution. Definition 4.1 defines the standard Markov property, which takes a DAG's d-separation statements and maps them to conditional independences. Compared to the S-Markov property from Definition 2.2, the Markov property only captures invariances present in a single distribution. However, in the complex real world, problems may be modeled with different distributions. For example, in machine learning, a common problem is generalizing learning to out-of-distribution settings.

[7] introduced a new characterization that extends the Markov property to account for experimental data arising from known-target interventions.

Definition 4.7 (I-Markov Property [7])**.** Consider the tuple of absolutely continuous probability distributions $(P_I)_{I\in\mathcal{I}}$ over a set of variables $\mathbf{V}$. A tuple $(P_I)_{I\in\mathcal{I}}$ satisfies the I-Markov property with respect to a graph $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ if the following holds for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:

(1) For $\mathbf{I} \in \mathcal{I}$: $P_I(y|w, z) = P_I(y|w)$ if $(Y \perp Z|W)_G$.

(2) For $\mathbf{I}, \mathbf{J} \in \mathcal{I}$: $P_I(y|w) = P_J(y|w)$ if $(Y \perp K|W\backslash W_k)_{G_{\underline{W_k},\overline{R(W)}}}$

Remark 1. We see that the I-Markov property fixes the intervention targets, $\mathbf{I} \in \mathcal{I}$ and then allows the graphical structure to change fitting the Markov property with respect to a **tuple** of distributions now rather than a single distribution.

Similarly, the S-Markov property allows one to fix the intervention targets in the case of known-target interventions, but more importantly generalizes to the setting with different domains and unknown-targets at the same time.

Experimental data can come with either known-target interventions, where the targets are explicitly perturbed, or from unknown-target interventions, where one knows an intervention took place, but is unsure of what nodes it possibly affects. This resulted in the Ψ-Markov property.

Definition 4.8 (Ψ-Markov Property [6])**.** Let $G = (\mathbf{V} \cup \mathbf{L}, \mathbf{E})$ denote a causal graph, let $\mathbf{P}$ denote an ordered tuple of distributions and let $\mathcal{I}$ denote an ordered tuple of interventional targets such that $|\mathbf{P}| = |\mathcal{I}|$. Tuple $\mathbf{P}$ satisfies the Ψ-Markov property with respect to the pair $\langle G, \mathcal{I}\rangle$ if the following holds for disjoint $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$:

(1) For $\mathbf{I}_i \in \mathcal{I}$: $P_i(y|w, z) = P_i(y|w)$ if $(\mathbf{Y} \perp \mathbf{Z}|W)_G$

(2) For $\mathbf{I}_i, \mathbf{I}_j \in \mathcal{I}$: $P_i(y|w) = P_j(y|w)$ if $(\mathbf{Y} \perp \mathbf{K}|\mathbf{W}\backslash\mathbf{W_K})_{G_{\underline{W_K},\overline{R(W)}}}$

where $\mathbf{K} := \mathbf{I}_i \Delta \mathbf{I}_j$, $W_K := \mathbf{W} \cap \mathbf{K}$, $R := \mathbf{K}\backslash\mathbf{W_K}$ and $\mathbf{R}(\mathbf{W}) \subseteq \mathbf{R}$ are non-ancestors of $\mathbf{W}$ in G.

Compared to the S-Markov property, the Ψ-Markov property does not allow us to characterize invariances with known-target interventions. More importantly, the Ψ-Markov property does not characterize invariances for distributions that occur in different domains. As we show in Example 9, 4, characterizing interventions separately from domain-changes is an important distinction to make.

D.4 Comparisons with Other Works

In this section, we explicitly discuss some subtleties compared to work that could be also seen as learning over multiple domains. We survey a few works in the area of causal discovery that touch upon structure learning in the presence of multiple distributions of data. We illustrate similarities and differences via examples.

D.4.1 Single-domain interventions with known-targets: $\mathcal{I}$-FCI [7, 44]

[44] characterize a MEC under interventions with known-targets. [7] further refines the characterization and shows an improved EC and learning algorithm, the $\mathcal{I}$-FCI algorithm. As shown in Ex. 9 and related simulation experiment in Experimental Results - Simulations, the S-FCI not only learns additional details when possible, but also does not learn incorrect statements compared to the $\mathcal{I}$-FCI algorithm.

Furthermore, in other works, such as [44] have made this assumption since observational data is typically cheaper and easier to collect compared to experimental data. However, it is also plausible that many times only experimental data is available and observational data is riddled with selection biases. For example, in a controlled experiment in the lab, one can control biological samples, but if collecting biological samples from an observational hospital setting, then the samples may contain selection bias depending on what sort of patients the hospital specializes in treating (e.g. cancer patients, or pediatric patients).

Another important connection with the work of $\mathcal{I}$-FCI is the unsoundeness of the orientation rules in the presence of known-target interventions. In F-FCI, R9 orienting inducing paths works only if $|S_k| = 1$, where $S_k = \mathbf{I}\Delta\mathbf{J}$ is the symmetric difference between different interventions. However, a trivial example where this is incomplete is given in Figure S2, where there is observational data and a joint intervention on $\{X, Y\}$. R9 of $\mathcal{I}$-FCI does not apply in this setting and thus one does not orient the edge between Y and Z. However, we should be able to deduce that $Y \to Z$. Although this example is somewhat trivial, the issue extends whenever $|S_k| > 1$ and this leaves room for improvement in the presence of known-target interventions. Providing a complete orientation rule such that $\mathcal{I}$-FCI and S-FCI can be complete in the presence of known-target interventions would be interesting future work.

On the other hand, consider the example shown in Figure S3, which shows an example, where one might try to orient the inducing path, but this would be incorrect.

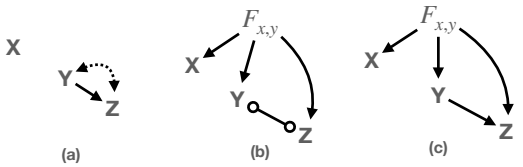

Figure S2: **Counter-example demonstrating that orientation rules for known-targets stemming from $\mathcal{I}$-FCI are not complete.** (a) shows the ground-truth graph. Given $\mathbf{\Psi} = \langle\{\}^1, \{X, Y\}^1\rangle$ intervention targets with $\mathcal{K} = [1, 1]$, (b) shows the graph learned using $\mathcal{I}$-FCI (or S-FCI). (c) shows what we should be able to learn due to the inducing path between $F_{x,y}$ and Z through Y.

D.4.2 Single-domain interventions with unknown-targets: Ψ-FCI [6]

[6] generalize the work of the $\mathcal{I}$-FCI and its EC characterization to the setting with unknown intervention targets and the authors propose a constraint-based learning algorithm for learning a Ψ-MEC, the Ψ-FCI algorithm. Given the results from Corr. 5, one might suspect that the S-FCI algorithm and the S-Markov characterization is just a relabeling of the Ψ-FCI and Ψ-Markov characterization.

Here, we construct an example demonstrating that when considering the domain setting, one can learn more than just naively applying the Ψ-FCI algorithm. Moreover, this demonstrates that the S-Markov characterization is a more refined EC characterization.

Example 10. Consider the selection diagram in Figure S4(a). The S-node pointing to Z indicates that there is a possible change in mechanism going from domain 1 to domain 2. Let $\mathbf{\Pi} = \{\Pi^1, \Pi^2\}$, $\mathbf{\Psi}^{\Pi} = \langle\{\}^1, \{X\}^1, \{\}^2\rangle$, $\mathbf{P}^{\Pi} = \langle P_1^1, P_2^1, P_1^2\rangle$ and $\mathcal{K} = [1, 0, 1]$. Assume we have access to an oracle to query for d-separation.

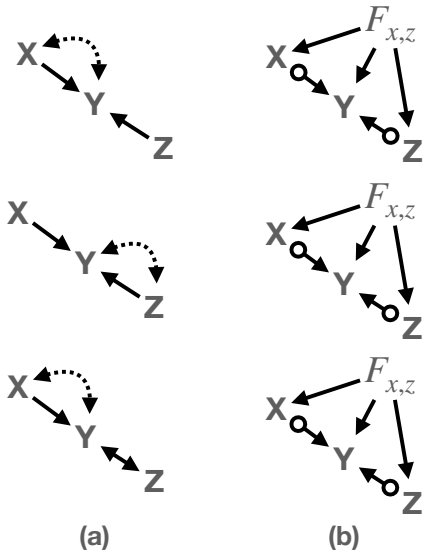

Figure S3: **Counter-example demonstrating orientation rules for known-targets stemming from $\mathcal{I}$-FCI are not complete.** (a) shows three different causal graphs. Given $\mathbf{\Psi} = \langle \{\}^1, \{X, Z\}^1 \rangle$, with $\mathcal{K} = [1, 1]$, we would learn the I-PAG in the column (b). The graphs in column (b) are all in the same MEC. The inducing paths for example from $Z \to Y$ cannot be oriented in this case, otherwise the edges would not be sound.

If one runs the Ψ-FCI algorithm, then there is no notion of multiple domains in the Ψ-Markov characterization. Therefore, we would ignore the domain superscript, and combine the two observational datasets. Running the algorithm results in the Ψ-PAG in Figure S4(b). Observe that the skeleton of the variables $\{X, Y, Z\}$ is correct. However, no orientations are learned. In contrast, Figure S4(c) shows the results of running the S-FCI algorithm. Observe that there is not only improved orientation by learning that $Y \to Z$, but also the augmented nodes provide additionally rich structure. For example, the S-PAG indicates that the only S-node present in the true selection diagram is one that points to Z. □

This example demonstrates that the characterization and S-FCI algorithm proposed in this paper improves upon the work of [6]. Note that we demonstrate subtle differences that show we improve upon the Ψ-FCI algorithm. The appendix of [6] also shows similar examples that illustrate subtle differences of the Ψ-FCI algorithm with respect to other works, such as [13, 14, 44, 71].

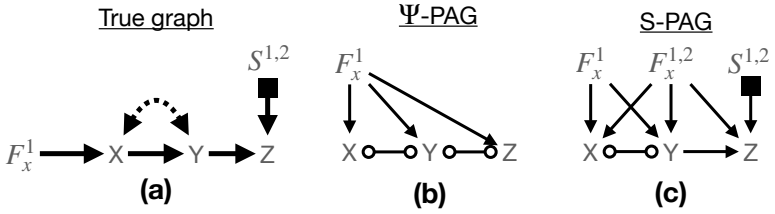

Figure S4: augmented graph representing an intervention and a S-node over domains 1 and 2 (a), the resulting Ψ-PAG one can learn using Ψ-FCI (b), and the resulting S-PAG one can learn using S-FCI (c).

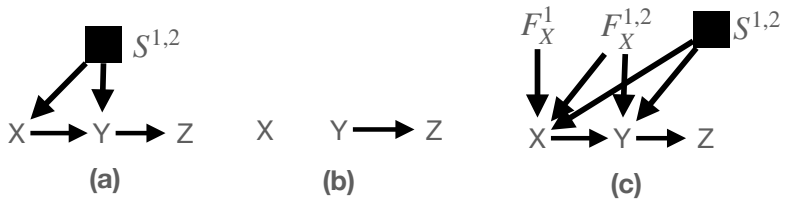

Figure S5: **Comparison of ICP vs SFCI** given ground-truth graph (a). Assume that we are given interventions $\mathbf{\Psi} = \langle \{\}^1, \{X\}^1, \{\}^2 \rangle$ and their corresponding distributions with known-targets $\mathcal{K} = [1, 0, 1]$. (b) is the graph learned by ICP. (c) is the S-PAG learned by S-FCI.

D.4.3 Invariant Causal Prediction [13, 35]

Invariant causal prediction (ICP) can identify the causal parents of a target variable under the assumption that the target's causal mechanism is invariant across environments [13, 35]. The work in [13] treats interventions and different regimes (i.e. domains) as similar concepts, whereas in this work, as explicitly noted by the S-Markov property, they are in fact quite different in subtle ways.

The work proposes for a target variable Y, to identify subsets S such that $P_i(Y|S)$ are invariant across all distributions $P_i(.)$ for all "environments" i. Interestingly, the paper provides sufficient conditions for the ICP framework to uniquely identify the true causal parents of Y. However, this requires the assumption of linear SCMs, the absence of latent confounders, and certain constraints on the set of interventions. It is interesting future work to explore the ideas introduced in this paper in the context of functional assumptions on the causal structure. However, we contrast our approach mainly with the idea of leveraging invariances across distributions.

Mainly, the authors in [13] suggest looking for invariances that hold across *all* domains, whereas we look for invariances across pairs of domains. Moreover, we also leverage different pairs of distributions to learn different information. For example, comparing interventions across domains allows one to learn invariances with respect to both the domain change and the intervention set. However, comparing interventions within a domain allows one to learn invariances with respect to the intervention set, with the implicit assumption that there are no other changes induced by a changing environment.

As an example, consider the graph and setting shown in Figure S5(a). We have known-target interventions $\mathbf{\Psi} = \langle \{\}^1, \{\}^2, \{X\}^1 \rangle$, $\mathcal{K} = [1, 1, 1]$ and their associated distributions $\mathbf{P}$. When applying ICP, one would recursively say discover parents and say we start with Z. Across all distributions, one would see that $P(z|y)$ is invariant, and thus $Y \to Z$. However, say one moves to the node Y next. There is no invariance for $P(Y|S)$ across all distributions, since for example the domain-shift from domain 1 to 2 through the S-node, $S^{1,2}$ affects Y. Thus, ICP may learn the graph in Figure S5(b). On the other hand, Figure S5(c) show the result of applying S-FCI and even the S-node structure is recovered.

D.4.4 Causal Discovery with Joint Causal Inference [14]

The work in [14] proposed "Joint Causal Inference" (JCI) as a framework that pools multiple datasets/distributions with unknown interventional targets and then employs a standard causal discovery algorithm to learn the causal graph, such as FCI. Namely, FCI-JCI is an adaptation of the FCI algorithm that learns causal graphs over the pooled datasets, combining different observational and interventional datasets. In [6] Appendix Section D.2, it is shown that Ψ-FCI explicitly can learn more than the JCI procedure. Moreover, [6] Appendix Section D.2 Proposition 6 demonstrates a proof that this holds in general for settings with at least three distributions. The basic intuition is that JCI compares everything relative to the observational distribution, which can miss invariances. On the other hand, comparing every pair of distributions is important for characterizing all possible invariances. Since JCI is already shown to characterize and learn less in a single-domain setting, the same will hold when we consider the multi-domain setting. We direct the readers to [6] for additional discussion on the single-domain setting comparing Ψ-FCI to JCI.

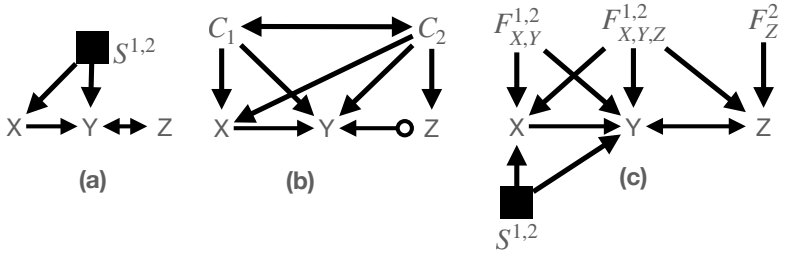

Figure S6: **Comparison of JCI vs SFCI** given ground-truth graph (a). Assume that we are given interventions $\mathbf{\Psi} = \langle \{\}^1, \{X, Y\}^2, \{X, Y, Z\}^2 \rangle$ and their corresponding distributions with known-targets $\mathcal{K} = [1, 1, 1]$. (b) is the graph is the S-PAG learned by S-FCI and (c) is the graph learned by JCI. The results hold if the interventions are unknown-targets as well, and even if there was observational data in both domains.

The pooling procedure constructs auxiliary context variables, $\mathbf{C} = \{C_i\}_{i=1}^{M}$ given datasets $\langle D_0, ..., D_M \rangle$, where D_0 corresponds to the "observational distribution and D_i corresponds to an "interventional" distribution. The algorithm pools the datasets into one, D^* and then appends context variables such that $\mathbf{C} = 0$ for D_0 and $C_i = 1$ if the sample corresponds to D_i, else $C_i = 0$. Thus there are an additional M columns in the dataset, which result in added nodes to the causal graph. When context nodes are added, $C_i \leftrightarrow C_j$ for all i, j and then $C_i \rightarrow V_j$ if there is a dependency among the C_i variable and the V_j variable.

In Figure S6, SFCI is shown to learn more than JCI. JCI learns the graph in Figure S6(b). SFCI learns the graph in (c) and importantly also estimates the S-node structure.

D.4.5 Causal Discovery with Nonstationary Changes [11, 36]

[11, 36] also uses auxiliary random variables to capture mechanism changes. JCI can be seen as an extension of this idea. Similarly to JCI, our approach differs in how we treat these auxiliary nodes and characterize the pairwise-distribution invariances in a more complete manner.

D.4.6 Multi-Domain Causal Structure Learning in Linear Systems [37]

The paper [37] proposes a causal discovery method that accounts for observations across multiple domains. However, the setting relies on the absence of latent confounders and also linearity in the SCM. In this work, we characterize the EC in the semi-Markovian and nonparametric setting.

D.5 Experimental Results - Simulations

All experiments are reproducible using the algorithm implementations at https://github.com/py-why/dodiscover and https://github.com/py-why/pywhy-graphs [72, 73].

D.5.0.1 Chain-Graph Experiment In this section, we demonstrate empirically through computational experiments that S-FCI learns more, or more accurate graphs relative to the true selection diagram.

In the first simulation, a very simple setup is done to confirm the presentation of Ex. 9. In this example, $G = \{Y \rightarrow X, S^{1,2} \rightarrow X\}$ is the selection diagram with the augmented-selection diagram shown in Figure S7(c). The ground-truth causal diagram and augmented graph are shown in Figure S7(a-b), neither of which encode the change in domain.

Data is generated using a linear SCM, where nodes have exogenous noise generated from a Gaussian distribution (μ, σ) where μ is generated uniformly in [-5, 5] and σ is generated uniformly in [0.01, 1.5], and edge weights are generated uniformly in [-5, 5]. Each node is a linear combination of its parents, where edge functions are generated uniformly from the following choices with "x" as the input: linear (x), quadratic (x^2), sin (sin(x)), or negative (-x). We repeat the experiment 10 times with sample sizes ranging from 500 to 5000 linearly spaced. At each parametrization, we repeat the experiment 5 times. We simulate two different domains, with the S-node pointing to Y indicating a possible change in mechanism between domain 1 and 2. In total, we generate two

different distributions, $\mathbf{P}^{\Pi} = \langle P_1^1, P_2^2 \rangle$, one per domain. Each distribution is interventional. We simulate a soft intervention on the node X by additively perturbing the values of X. We encode different soft interventions in domain 1 and 2 (i.e. the mechanisms have the same target, but different mechanisms; $\mathbf{\Psi}^{\Pi} = \langle \{X^a\}^1, \{X^b\}^2 \rangle$). We assume the targets are known, $\mathcal{K} = [1, 1]$.

Using the ground-truth diagram as an oracle for conditional independence and conditional invariance testing (of the form listed in Def. 2.2), we can get different ECs, which are shown in Figure S7(d-f). As we expect from Ex. 9, the $\mathcal{I}$-FCI arrives at the incorrect causal conclusion, $X \to Y$. Next, using partial correlation and the Kernel conditional discrepancy test [74], we test this setting with finite data. In Figure S8, we see that it is always the case that S-FCI learns the correct graph even with finite data.

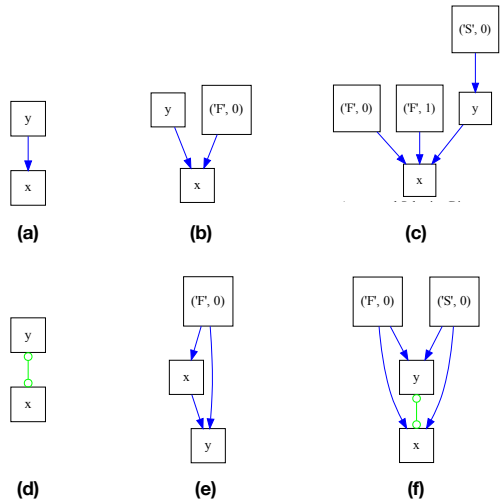

Figure S7: Comparing S-FCI vs FCI vs $\mathcal{I}$-FCI in a simulation with two known-target interventions with different mechanisms on X - The top row shows the true diagrams: (a) is the true causal diagram, (b) is the augmented diagram encoding the intervention on X, (c) is the augmented graph that shows the interventions on X in each domain and the S-node indicating a possible change in mechanism for Y. The bottom row shows the learned EC with an oracle for querying d-separation - (d) the PAG learned by the FCI algorithm, (e) the I-PAG learned by the $\mathcal{I}$-FCI algorithm and (f) the S-PAG learned by the S-FCI algorithm.

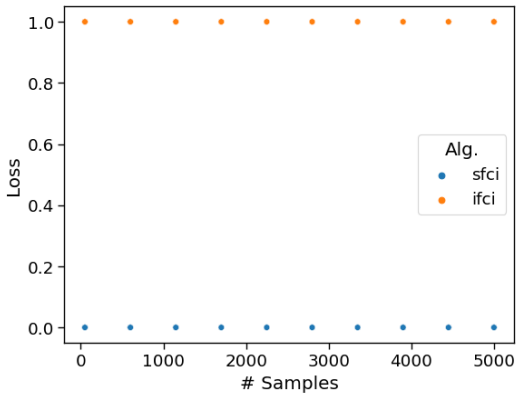

Figure S8: S-FCI learns the correct orientation consistently given known target interventions in multiple domains compared to $\mathcal{I}$-FCI in linear SCMs following the two-node setting.

D.5.1 Analysis of Protein Sequencing

As motivated by Ex. 4 and 9, we next analyze a protein sequencing Sachs dataset [22], where different perturbations of proteins were made, and then responses from other proteins were observed. The

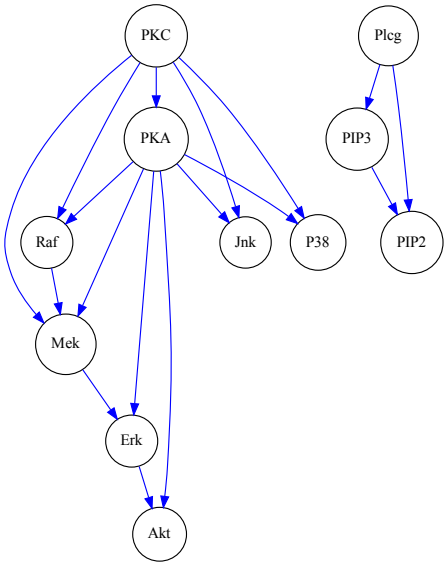

Figure S9: The presumed ground-truth graph for the protein experiments from [22]. Imported from bnlearn [76].

ground truth graph is given by [22] and is shown in Supplemental Figure S9. We utilize this dataset because it is a commonly used dataset to evaluate causal discovery in many papers [6, 7, 38, 75]. We run S-FCI and get the results shown in Figure S10, where various structures such as the cluster among (PIP3, PIP2, Plcg) is detected and certain orientations in the larger graph are also correctly detected. As a result, these two experiments provide a realistic setting in which S-FCI could plausibly be used [5].

D.5.2 Simulated Data

In this next section, we present some experiments validating that adding additional data across multiple domains improves upon the structure by helping orient additional edges.

The ground-truth graph is shown in Figure S11(a). We forward-sample discrete data according to the graphical model and implement categorical data with cardinality of "3" per node. We then sample a random conditional probability distribution (CPD) for each node in topological order using pgmpy [77]. By specifying the full conditional distributions for each node as a function of its parents, this now specifies the full SCM. We then proceed with four different settings:

1. From this SCM, we sample 30,000 samples to denote the observational distribution, obs. We will denote this SCM as coming from domain 1. We run FCI on the data and obtain Figure S11(b).
2. Next, we generate 30,000 samples of interventional data by intervening on the 'D' node, generating a new CPD for node D. Then we run the $\mathcal{I}$-FCI, or Ψ-FCI algorithm depending on if we assume the intervention is a known-target or not. Regardless of the algorithm, the graph learned is in Figure S11(c).
3. Next we generate a domain-shift that changes the distribution of node X and C. I.e. in the corresponding selection diagram of (a), this would have the additional edges $X \leftarrow S^{1,2} \rightarrow$

[5] Real world data with ground truth selection diagrams and observational and interventional data collected over multiple domains is a big challenge that is necessary to evaluate multi-domain causal discovery algorithms. This paper partially addresses this need by leveraging real single-domain data and using that data to generate plausible datasets to simulate the multi-domain setting. Additional research is needed that generates this dataset in the real world from experiments and observations.

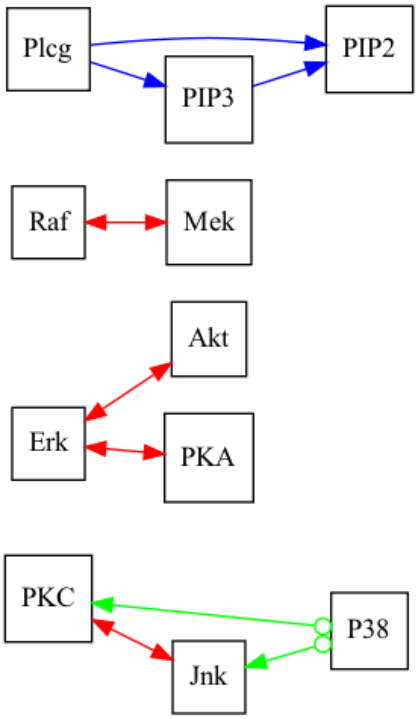

Figure S10: Shows the learned S-PAG of the Sachs dataset.

C. This generates a new SCM that represents domain 2. Combining the observational datasets from domain 1 and the domain 2, we can run FCI again and obtain the Figure S11(d).

4. We also simulate an intervention that occurs on node D again this time in domain 2. By pooling the interventional datasets and the observational datasets and naively ignoring the difference in domain, we can re-run the Ψ-FCI algorithm and obtain the graph in Figure S11(e).
5. Finally, taking all datasets together and applying the S-FCI algorithm, we obtain the result in Figure S11(f).

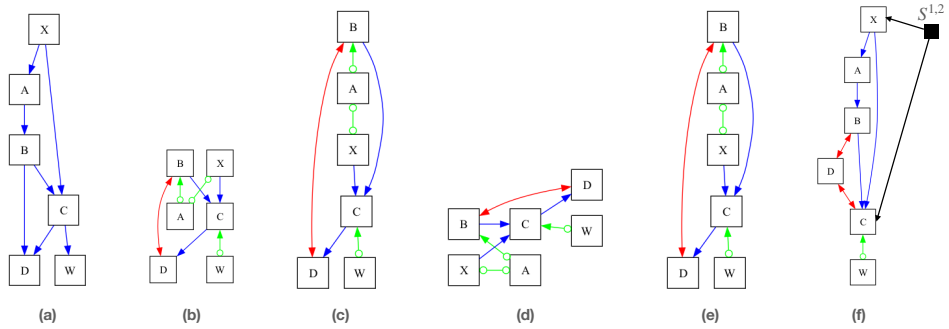

Figure S11: **Example simulation comparing FCI, Ψ-FCI and S-FCI using the same datasets** with ground-truth graph in (a). Running FCI on single-domain observational data results in (b). Running Ψ-FCI on single-domain observational and interventional data results in (c). Stacking the multi-domain observational data, ignoring the domains and running FCI results in (d). Stacking the multi-domain observational and interventional data, ignoring the domains and running Ψ-FCI results in (e). Running S-FCI on the same dataset as (e) without ignoring domains results in (f). (f) learns the most correct graph relative to ground-truth (a).

D.6 Discussion on Assumptions

In this paper, various assumptions are made. Here, we include additional discussion on the justification of those assumptions.

- (Shared causal structure assumption 2) - This assumption indicates that across different domains, there is no change in the input arguments of the functionals in the SCM. That is $x_i \leftarrow f(arg1, arg2, ..., argk)$ has the same input arguments for any f that is assigned in different SCMs corresponding to different domains. This limits the scope of functional changes across domains, but is realistic since many real-world applications potentially share causal structure even with a domain-shift. For example, in human brains when listening to speech, from subject to subject, it is reasonable to suspect that there is a domain change in how brain activity manifests. Although the amplitude and frequencies of brain activity may differ, it is reasonable to assume that the brain regions involved are the same from subject to subject. That is, the visual cortex in the occipital lobe is involved in visual processing, with Wernicke's area is involved in downstream speech processing. Thus a causal selection diagram with assumption 2 preserves the causal structure of speech processing, while allowing any other kinds to the functionals of the SCM.
- (Interventions have different mechanisms across domains) - This assumption specifies that any intervention cannot be exactly the same when they occur in different domains. Let $x_i \leftarrow f_i$ be the function that assigns the value for node x_i in the SCM before an intervention. This essentially restricts the sub-model from the intervened SCM to not be able to have $f_i \leftarrow f_i'$ for two interventions that occur in different domains. If the interventions occur in different domains, then they must always be of the form $f_i \leftarrow f_i'$ in domain k and $f_i \leftarrow f_i''$ in domain l, where $f_i' \neq f_i''$. Thus even if the same nodes are intervened in two different domain settings, there is still a difference in their distributions, which is encoded by the S-node edge. This assumption is realistic because if a scientist is making the assumption that there is say genomic sequencing data coming from two different domains (e.g. lab settings, or biological specimens), then even if the intervention is applied to the same gene, or set of genes, it is highly unlikely that the intervention has the exact same effect in the different domains. Hence, running the same sequencing experiemnts in different labs are known as batch effects and accounting for them allows one to make causal inferences across domains [29].
- (Soft interventions) - Currently, no characterization for a Markov class with respect to hard interventions exists. Each intervention alters the graph, making a consistent object for constraint collection challenging. An analogous Proposition 1 version is difficult due to intervention-induced graph changes. Hence, mapping invariances to augmented graph's d-separation differs. Our work aims to illuminate properties and subtleties for potential broader characterization later.

In addition to these assumptions, we make the assumption that any intervention does not reproduce a domain-shift. This assumption is implicit in a single-domain setting, where interventions are assumed to actually intervene on the system, such that there is a difference relative to the observational distribution. In the multi-domain setting, then these interventions perturb the distribution space, but is independent from a domain-shift and is assumed to never exactly equal an observational distribution in another domain. The reason for this assumption is related to faithfulness and Definition 2.2. For example, if we have two domains, and intervention $\{X\}^1$ in domain 1 and observation $\{\}^2$ in domain 2. The intervention takes place in domain 1. If it perfectly reproduced the domain-shift going from domain 1 to domain 2, then it would provide an additional invariance that is not encoded in the graph.

D.7 Background and Additional Preliminaries

In this section, we provide additional background notation and concepts relevant for the proofs and theoretical concepts introduced in this paper.

Additional Notation

A path *p* from X to Y in G is a sequence of distinct nodes $\langle X, ..., Y \rangle$ where each pair of consecutive nodes is adjacent in G. A directed path (also known as a causal path) from X to Y is a path where

all edges are directed $X \to \cdots \to Y$. A possibly directed path from X to Y is a path where no arrowhead is pointing to X. A star on edge endpoints is used as a wildcard to denote circle, arrowhead, or tail.

We say if $X \to Y$, then X is a parent of Y. If there is a (possibly) directed path from X to Y, then X is a (possible) ancestor of Y and Y is a (possible) descendant of X. The convention is that every node is also a descendant and ancestor of itself. The sets of parents and (possible) descendants of X in G are denoted by $Pa(X, G)$ (or just $Pa(X)$ when it is unambiguous) and (Poss)De(X, G) respectively. Similarly, we would also write $PossCh(X)$ as the possible children of X, and $NonDesc(X)$ as the definite non-descendants of X. A definite non-descendant, Z, is one where there is no possibility of Z being a descendant of X. This can occur if there is a arrow-endpoint ending at X, or a tail-endpoint ending at Z.

A triple $\langle X, Y, Z \rangle$ is an unshielded triple if X and Y are adjacent, Y and Z are adjacent, and X and Z are not adjacent. If both edges are into Y, then the triple is referred to as unshielded collider. A path between X and Y, $p = \langle X, ..., W, Z, Y \rangle$, is discriminating for Z if every node between X and Z is a collider on p and is a parent of Y. Two MAGs are Markov equivalent if and only if (1) they have the same adjacencies; (2) the same unshielded colliders; and (3) if a path p is a discriminating path for Z in both graphs, then Z is a collider on p in one graph if and only if it is a collider on p in the other. A PAG represents an MEC of a MAG and is learnable from data. The output of the celebrated FCI algorithm is a PAG, which is proven sound and complete for the corresponding MEC [39].

Structural Causal Models

We use Structural Causal Models (SCMs) [1] as our basic semantical framework. A SCM is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(u) \rangle$, where 1) $\mathbf{U}$ is a set of exogenous (latent) variables, 2) $\mathbf{V}$ is a set of endogenous (observed) variables, 3) $\mathbf{F}$ is a set of functions that determine the values of endogenous variables (i.e. $v \leftarrow f_V(\mathbf{pa}_V, \mathbf{u}_V)$ is a function with $\mathbf{pa}_V \subseteq \mathbf{V}$ $\{V\}$ and $\mathbf{U}_V \subseteq \mathbf{U}$ and 4) P(u) is a joint distribution over exogenous variables, $\mathbf{U}$.

Each SCM induces a causal diagram, G [52], where every variable $v \in \mathbf{V}$ is a vertex and directed edges in G correspond to functional relationships specified by $\mathbf{F}$ and bidirected edges represent common exogenous variables between two vertices. Within the structural semantics, an intervention by setting $X = x$ is represented with the do-operator, which encodes the operation of replacing the original functions of X (i.e. $f_X(\mathbf{pa}_X, \mathbf{u}_X)$) by the constant x and then induces a submodel M_x and corresponding interventional distribution $P(v|do(x))$.

Definition 4.9 (Selection Diagrams)**.** Let $\langle M, M^* \rangle$ be a pair of SCMs relative to the domains $\langle \pi, \pi^* \rangle$, sharing a causal diagram G. $\langle M, M^* \rangle$ is said to induce a **selection diagram** D, if D is constructed as follows: every edge in G is also an edge in D; D contains an extra edge $S_i \to V_i$ whenever there exists a discrepancy $f_i \neq f_i^*$, or $P(U_i) \neq P^*(U_i)$ between M and M^*.

Selection diagrams are causal graphs imbued with extra selection "S-nodes". Selection diagrams are induced from tuples of SCMs rather than a single one, since they represent different underlying SCMs.

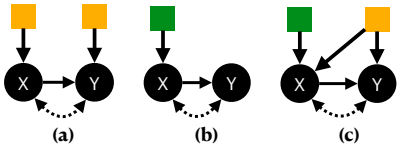

Figure S12: A selection diagram between source domain "i" and target (a), a selection diagram between source domain "j" (b) and target and a joint selection diagram (c).

Ancestral Graphs

A mixed graph can contain directed and bi-directed edges. A is an ancestor of B if there is a directed path from A to B. A is a spouse of B if $A \leftrightarrow B$ is present. If A is both a spouse and an ancestor of B, this creates an almost directed cycle. A mixed graph is ancestral if it does not contain directed or almost directed cycles. It is maximal if there is no inducing path (relative to the empty set) between

any two non-adjacent nodes. A Maximal Ancestral Graph (MAG) is a graph that is both ancestral and maximal [60]. Given a causal graph $G(V \cup L, E)$, a unique MAG M_D over V can be constructed such that both the independence and the ancestral relations among V are retained; see, [60]. Ancestral graphs, such as MAGs represent a Markov equivalence class (MEC) of a DAG, G. PAGs represent the unique MEC of a MAG. Therefore a DAG maps to a unique PAG, representing the MEC of the conditional independence statements. However, a PAG represents a class of different DAGs that encode the same conditional independence statements.

MAGs and PAGs are ECs of causal diagrams. In this paper, analogues, such as the S-MAG (Def. 2.4) and S-PAG (Def. 3.1) are introduced and relevant for encoding an EC of selection diagrams.

m-separation and m-connectedness

In this section, we briefly review the graphical criteria m-separation and m-connectedness, which is a generalization of d-separation and d-connectedness [60]. First, a necessary definition to fully understand and characterize m-separation is the notion of definite colliders and definite non-colliders in PAGs as these are a way to determine definite status along a path that does not need oriented end points.

Definition 4.10 (Definite collider and non-colliders)**.** Let $\langle A, B, C \rangle$ be a consecutive triple along a path p in a PAG, G. B is a definite collider on p if both edges are into B (i.e. arrowhead endpoint at B). B is a definite non-collider on p if either one of these is true:

i) One of the edges is out of B ($A \leftarrow B *\!\!-\!\!* C$, or $A *\!\!-\!\!* B \rightarrow C$. ii) Both edges have a circle-endpoint at B, and there is no edge between A and C (i.e. $\langle A, B, C \rangle$ is unshielded). This looks like $A *\!\!-\!\!o B o\!\!-\!\!* C$.

Otherwise B has a non-definite status along p.

A definite status path p between nodes X and Y is m-connecting given a set of nodes $\mathbf{Z}$ (with $X, Y \notin \mathbf{Z}$) if every definite non-collider on p is not in $\mathbf{Z}$ and every collider in p has a descendant in $\mathbf{Z}$. A possibly m-connecting path between X and Y given $\mathbf{Z}$ is a path where every definite non-collider on the path is not in $\mathbf{Z}$ and every collider has a possible-descendant in $\mathbf{Z}$.

If $\mathbf{Z}$ blocks all definite status paths between X and Y, we say that X and Y are m-separated given $\mathbf{Z}$. Otherwise X and Y are m-connected. If $\mathbf{Z}$ blocks all possibly m-connecting paths between X and Y, we say that X and Y are $\hat{m}$-separated given $\mathbf{Z}$.

D.8 Additional Example Illustrating S-FCI Subtleties

We illustrate a few additional examples that augment Ex. 9 to show how observational data is necessary to differentiate interventions and domain-associated mechanism changes.

Example 11. Consider the same setup as in Example 9 with the ground-truth selection diagram in Figure 3. This time, in Π^1, say we have access to also observational distributions. Specifically, $\mathbf{\Psi}^{\Pi} = \langle \{Y\}^1, \{\}^1, \{\}^2 \rangle$ with $\mathcal{K} = [1, 1, 1]$ and $\mathbf{P} = \langle P^1_1, P^1_2, P^1_3 \rangle$, where we have the ability to experiment on Y proteins and collect passive observations in a laboratory setting, and we have access to observational data in a hospital setting.

Since we have access to the observational dataset in domain Π^1, then it would be possible to learn the correct orientation of $X \leftarrow Y$. This is because within the same domain, the observational and interventional distributions can be compared.

Furthermore, we know that a potential selection diagram that with the correct cross-domain invariances is one with the S-node, $Y \leftarrow S^{1,2} \rightarrow X$ from the F-node that is constructed comparing $\{\}^1, \{\}^2 \in \mathbf{\Psi}$, $F^{1,2}_{\{\}}$. However, note regardless that one cannot learn that the S-node only points to X. This is because of the inducing path between $S^{1,2}$ and Y through X that makes the two nodes adjacent in all graphs of the EC.

Note, the same result could be obtained if one had an interventional dataset that simultaneously intervenes on $\{X, Y\}$ because the symmetric difference between $\{X, Y\}^1 \Delta \{Y\}^1 = \{X\}$ results in an F-node, F^1_X associated with the known-target X. The F-node would be adjacent to both X and Y, and through R9' of the S-FCI algorithm, we could orient the edge $X \rightarrow Y$ correctly. □

In this example, we demonstrate it is very valuable to obtain the observational distribution in separate domains to calibrate against, which allows one to differentiate interventional effects from change-in-domain effects. This next example demonstrate that nothing is learnable in the degenerate case where no data is given in one of the domains.

Example 12. Consider again the setting in Example 9. This time however, we do not have access to observational data in Π^* (i.e. we do not have experimental, or observational data in the hospital setting). The only thing we can learn is whether or not there is an adjacency between X and Y given the data in Π^1. However, at this point, we do not know and cannot learn if there are S-nodes pointing to X, Y, both, or neither because we simply do not have access to data in one of the domains. It is impossible to know if there are conditional invariances when moving from the lab to the hospital that might affect the protein level expressions without some type of data in the hospital setting to calibrate against. Experimental, or observational data in the relevant "target" domain is necessary for causal discovery in the multi-domain setting. □

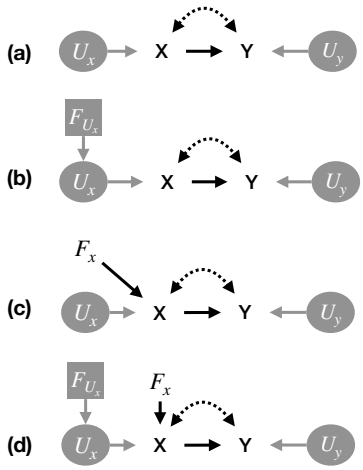

Figure S13: Demonstrating how nodes represent exogenous variables (a), mechanism changes (e.g. S-nodes) (b), interventions (c) are encoded in a causal diagram and then (d) an augmented graph, where interventions and change in domain are represented. The node in (b) is represented as a square node, and can be viewed as an "intervention" on the exogenous variables. In a selection diagram, F_{U_x} can be a S-node pointing to X. Nodes in gray are exogenous.

This example illustrates that it is not possible to learn relevant multi-domain graphical structure without data present in all domains.

Although Figure S13 suggests in some way that observations from multiple domains and interventions within a single domain can be viewed similarly, the next example demonstrates that not accounting for the domains, when data is collected from multiple domains can lead to incorrect characterization. This incorrect characterization is precisely what leads to the incorrect result of the $\mathcal{I}$-FCI algorithm in Example 9. Why is the setting where observational data not present more subtle? We illustrate how the characterization of the Markov property when comparing interventions across domains without observational data can lead to incorrect characterizations under our assumptions.

Example 13 (I-Markov vs S-Markov property)**.** Let G be a causal diagram as shown in Figure 1(a). Let $\mathbf{\Pi} = \langle \Pi^1, \Pi^2 \rangle$ be the set of domains representing the lab (Π^1) and the hospital (Π^2) and assume we have access to data where proteins are perturbed in the lab and the intent is to utilize that information along with observational data in the hospital (e.g. [22]). These are a tuple of distributions $\mathbf{P} = \langle P_1^1, P_1^* \rangle$ with intervention targets $\mathbf{\Psi}^{\Pi} = \langle \{Y\}^1, \{\}^* \rangle$ and $\mathcal{K} = [1, 1]$, where X represents some protein in the dataset.

Given the known-target intervention and observation, one can analyze the I-Markov property. Due to $(X \perp Y)_{G_{\overline{Y}}}$, would encode the additional constraint $P_1^1(X) = P_1^*(X)$ thus satisfying the I-Markov property. However, that is not true, since $\mathbf{P}$ is generated from the diagram with a S-node $S_x^{1,*}$ changing the mechanism of X. The S-Markov property constraint correctly encodes this because even though $(X \perp Y)_{G_{\overline{Y}}}$, we have $X \not\perp S^{1,*})_G$, thus the invariance is not valid. □

Even though the intervention target is known, we see that the I-Markov property does not correctly encode distributional constraints. We will explore this in more details in the next few sections.

S-nodes can be seen as analogous to interventions with the exception that S-nodes occur by some unknown change of distribution in either the exogenous variable distribution, U, or the function of the endogenous variable the S-node is pointing to, f_i. Consider a simple causal bow-graph. Figure S13(a) shows how this can be represented with its exogenous variables explicitly shown. In general, we never know the true structure of the exogenous variables since the true causal model is unknown. In Figure S13(b), an S-node pointing to X could be viewed as a soft-intervention on U. On the other hand, in Figure S13(c), soft-interventions are represented by F-nodes in an augmented graph that change the distribution of X [7, 54, 57].

Now, if we look at Figure S13b and c, then we notice that the conditional independence statements are the same. Notice that $S_x \not\perp Y|X$ and $F_x \not\perp Y|X$, and thus the soft-intervention is graphically similar to the S-node because we do not observe $U_x \in U$. Despite their similarities, it is important to explicitly note some subtle differences. As demonstrated in Ex. 4, the distribution changes between domains as denoted by a S-node is always present, whether the data is collected in an observational, or experimental setting. Thus S-nodes and F-nodes carry different meanings. However, in the purely observational setting, they are in some sense equivalent.

D.9 Broader Impact and Forward Looking Statements

The development of new causal discovery algorithms has the potential to improve our understanding of complex systems, and to help identify the causal factors underlying important societal issues. By improving our ability to learn causal relationships from observational and interventional data across multiple domains, your work could ultimately lead to more effective interventions to address these issues that are transportable across operating domains. Beyond the causal inference community, we expect that our results will enable fundamental contributions in various fields, including biology [22], epidemiology [65], economics [78] and neuroscience [21].

One significant research direction is to study how to relax the assumption that the joint selection diagram does not contain structural differences among the different domains. Additionally, it will be important in future research to develop new benchmarks that reflect this emerging multi-domain causal discovery paradigm to evaluate algorithms. Another important research question is how to perform transportability inference within this newly introduced equivalence class. Transportability of causal effects, also known as "external validity" [79, 80], "meta-analysis" [81], "quasi-experiments" [82], "heterogeneity" [83], is a critical task that has been studied under the assumption that a well-specified selection diagram is available. It will be important to develop algorithms for transportability inference given the selection diagrams' EC and develop algorithms for computing causal effects from an EC of selection diagrams. This would enable scientists to perform completely data-driven causal analysis across multiple domains.

D.10 S-FCI Algorithm Additional Details

Here, we expand on the S-FCI algorithm and its details. The inner-workings of the S-FCI algorithm are introduced in Algorithm 1. Here, we provide details for the rest of the algorithm.

D.10.1 S-FCI Algorithm Details

D.10.1.1 Creating augmented graph Alg. D.2 describes how the augmented graph is created by adding nodes that map to pairs of distributions, and optionally symmetric difference targets.

Algorithm D.2 Generalized Augmenting Nodes - $\mathbf{S}$ is the set of S-nodes, $\mathcal{F}^{\Pi}$ is the set of F-nodes over each domain, $\mathcal{K}$ is the vector of known intervention targets, $\mathbf{H}$ is the set of intervention targets mapping each pair of known-target interventions, σ is the mapping of each pair of distributions within each domain and $\mathbf{V}$ is the set of nodes in the graph.

function CREATEAUGMENTEDNODES($\mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V}$)
  $(\mathbf{S} = \phi, \mathcal{F}^{\Pi} = \phi, \mathbf{H} = \phi, \sigma : \mathbf{N} \times \mathbf{N} \to 2^V \times 2^V$
  $k \leftarrow 0$
  (Comparing distributions: Add F-nodes and S-nodes)
  for all pairs $\mathbf{I}_l^i, \mathbf{J}_m^j \in \mathbf{\Psi}^{\Pi}$ **do**
    $k \leftarrow k + 1$
    if $\mathbf{I}_l^i = \{\}^i, \mathbf{J}_m^j = \{\}^j, i \neq j$ **then**
      Add S^{ij} to $\mathbf{S}$
    else
      Add F_k^{ij} to $\mathcal{F}^{\Pi}$
      if $\mathbf{I}_l^i$ and $\mathbf{J}_m^j$ are known-targets and $i = j$ **then**
        $H_k^{ij} = \mathbf{I}^i \Delta \mathbf{J}^j$
        Add H_k^{ij} to $\mathbf{H}$
    $\sigma(k) = (l, m)$: **(Maps the kth F-node to distributions l and m)**
  return $\mathbf{S}, \mathcal{F}^{\Pi}, \mathbf{H}, \sigma$

D.10.1.2 Generalized Multi-Domain Skeleton Discovery Alg. D.3 describes a generalized algorithm for performing constraint-based skeleton discovery, which allows our algorithm to choose a method for choosing candidate conditioning sets, $CondSel$. For example, one may use all possible combinations of nodes (e.g. the SGS algorithm does this [40]), or the neighbors of the nodes (e.g. the PC algorithm does this [5]), or the possibly d-separating sets (e.g. in RFCI algorithm [42]). Alg. D.4 describes how to infer the skeleton structure using constraints found in the data. For instance, the first else-if statement states that all F-nodes are by construction separated. The second else-if statement states that an F-node will be separated from another node given a specific kind of invariance described in Condition 2 of Def. 2.2.

Algorithm D.3 Generalized Skeleton Discovery - G is the augmented causal diagram from Def. 2.3, $CondSel$ is the conditioning selection function for determining how to select candidate separating sets Z, P_{max} is a hyperparameter controlling the maximum size of the conditioning set

function GENERALIZEDSKELETONDISCOVERY($G, CondSel, P_{max}$)
  $G = (V \cup \mathcal{F}, E \cup E_{\mathcal{F}})$
  while $p < P_{max}$ **do**
    for $X \in V$ **do**
      for $Z \in CondSel(X, p)$ **do**
        if $(X \in \mathcal{F} \cap Y \in \mathcal{F})$ **then**
          $SepSet(X, Y) \leftarrow \phi, Sep(X, Y) = True$
        else
          $(SepSet(X, Y), Sep(X, Y) \leftarrow$ Generalized Do-constraints (see Alg. D.4)
        if Sep(X,Y) = True **then**
          Remove (X, Y) edge in graph G

Algorithm D.4 Generalized Do-Constraints - $\mathbf{\Psi}^{\mathbf{\Pi}}$ is the intervention targets per N domains, $\mathbf{\Pi}$; $\mathcal{K}$ are the known targets; $\mathbf{V}$ are the relevant causal variables.

function GENERALIZEDDOCONSTRAINTS($X, Y, \mathbf{S}, \mathcal{F}^{\Pi}, \sigma, \mathbf{\Psi}^{\Pi}, \mathcal{K}, \mathbf{V}$)
  $(\mathcal{F}^{\Pi} = \phi, SepSet = \phi, \sigma : \mathbf{N} \to 2^V \times 2^V$
  $\mathbf{V} \leftarrow \mathbf{V} \cup \mathcal{F}^{\Pi}$
  if $X, Y \notin \mathcal{F}^{\Pi}$, and $X, Y \notin \mathbf{S}$ **then**
    for $I^i \in \mathbf{\Psi}^{\Pi}$ **do**
      for $W \subseteq V \backslash \mathcal{F}$ **do**
        if $P_I^i(y|w, x) = P_I^i(y|w)$ **then**
          $SepSet = W \cup \mathcal{F}^{\Pi} \cup \mathbf{S}$
          SepFlag=True
  else if $X \in \mathbf{S}, Y \in V$ **then**
    $(l, m) \leftarrow \sigma(k)$
    for $W \subseteq V \backslash \mathcal{F}$ **do**
      if $P_l^i(y|w, x) = P_m^j(y|w)$ **then**
        $SepSet = W \cup \mathcal{F}^{\Pi} \cup \mathbf{S} \backslash S^{i,j}$
        SepFlag=True
  else if $X, Y \in \mathcal{F}^{\Pi}$ **then** (X and Y are both F-nodes)
    SepSet = $\mathcal{F}^{\Pi} \cup \mathbf{S} \backslash \{X, Y\}$
    SepFlag=True
  else if $X, Y \in \mathbf{S}$ **then** (X and Y are both S-nodes)
    SepSet = $\mathcal{F}^{\Pi} \cup \mathbf{S} \backslash \{X, Y\}$
    SepFlag=True
  else if ($X \in \mathcal{F}^{\Pi}$ and ($Y \in V$), so let $F_k^{i,j}$ denote X (X is a F-node representing a distribution between domains i and j, and Y is a normal node in $\mathbf{V}$) **then**
    $(l, m) \leftarrow \sigma(k)$
    for $W \subseteq V \backslash \mathcal{F}$ **do**
      if $P_l^i(y|w, x) = P_m^j(y|w)$ **then**
        $SepSet = W \cup \mathcal{F}^{\Pi} \backslash \{F_k^{i,j}\} \cup \mathbf{S}$
        SepFlag=True

D.10.1.3 Generalized Multi-Domain Orientation Rules

- We restate the orientation rules presented in 3 for completeness of the appendix.

Algorithm D.5 Generalized Orientation Rules - G is the causal diagram, $SepSet$ are the separating sets that were learned, $\mathcal{F}^{\Pi}$ is the set of F-nodes, H is the set of known-intervention targets and S are the S-nodes.

For every unshielded triple (X, Y, Z), if $Z \notin SepSet(X, Y)$ orient it as $X *\!\!\rightarrow Y \leftarrow\!\!* Z$
Phase IIb: Apply logical orientation rules
R1-7: Apply 7 FCI rules from [39] and following two rules until none apply.
Rule 8': For any $F_k^{i,j} \in \mathcal{F}^{\Pi}$, orient adjacent edges out of $F_k^{i,j}$.
Rule 9': For any $F_k^{i,j} \in \mathcal{F}^{\Pi}$, that is adjacent to a node $Y \notin H_k^{i,j}$, if $i = j$ and $X \in H_k^{i,j}$ and $|H_k^{i,j}| = 1$, orient $X \to Y$.